Universidad Autónoma de Madrid
Escuela politécnica superior

Doble Grado en Ingeniería Informática y Matemáticas

Trabajo Fin de Grado

TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS DE CALIDAD DEL AIRE

Autor: Carlos Ruiz Pastor
Tutor: Irene Rodríguez Luján
Ponente: José Ramón Dorronsoro Ibero

MAYO 2017

TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS DE CALIDAD DEL AIRE

Autor: Carlos Ruiz Pastor
Tutor: Irene Rodríguez Luján
Ponente: José Ramón Dorronsoro Ibero

Departamento de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
MAYO 2017

Resumen

La contaminación del aire supone, cada vez más, una amenaza a la salud humana y una preocupación medioambiental para el futuro. En concreto el dióxido de nitrógeno (NO_2), generado principalmente por el tráfico en zonas urbanas, alcanza en la actualidad niveles peligrosos para la salud. Los gobiernos de diversos países y ciudades han propuesto medidas para reducir esta contaminación. Es destacable el caso de la ciudad de Madrid, obligada a cumplir la normativa europea que impone unos niveles máximos de NO_2 en las ciudades. Para hacer una mejor gestión de las medidas, que implican limitaciones en términos de velocidad máxima en algunas vías o incluso restricciones de acceso y/o aparcamientos en el centro de la ciudad, es especialmente interesante la capacidad de obtener predicciones precisas de la concentración de NO_2 en la ciudad. En esta labor juega un papel fundamental el aprendizaje automático, una rama de la inteligencia artificial encargada del diseño de modelos de predicción, que son esenciales para anticiparse a los niveles de contaminación.

Por tanto, este trabajo tiene el objetivo de construir un modelo predictivo de la concentración de NO_2. Para ello se utilizan datos históricos de la concentración de NO_2 en la ciudad, datos reales del tráfico y predicciones de la metereología en Madrid. Previamente a la construcción del modelo, se han analizado las características de los datos utilizados así como de sus fuentes. El Ayuntamiento de Madrid, en una iniciativa por promover la ciencia de datos, ofrece de forma gratuita la información del tráfico y la calidad del aire en la ciudad, obtenida a través de estaciones y sensores colocados en distintas zonas. Los datos de predicciones metereológicas se obtienen de forma gratuita de la Administración Nacional Oceánica y Atmosférica (NOAA) de Estados Unidos.

En este primer estudio se restringe el modelo a cuatro estaciones de calidad del aire ubicadas en Plaza de España, Escuelas Aguirra, Méndez Álvaro y Plaza del Carmen, utilizando la carga de tráfico en los puntos de medida más cercanos. De las predicciones metereológicas se utilizan variables de temperatura, presión, velocidad del viento y humedad relativa. A partir de esta información se puede predecir la concentración de NO_2 con relativo éxito. Utilizando una regresión lineal con la regularización de ElasticNet se obtienen predicciones con un error relativo medio (MRE o MAPE) del 20 %, se ha comprobado que los modelos que incluyen las variables del tráfico mejoran a los que únicamente se basan en los datos históricos del NO_2, mientras que la inclusión de datos de predicciones meterelógicas no es tan relevante en los resultados. Además se ha implementado un modelo de clasificación de alertas que, utilizando el algoritmo XGBoost, consigue predecir correctamente la mayoría de las alertas.

Los resultados obtenidos en este TFG son prometedores y cabe esperar que la inclusión de nuevas variables y/o el uso de otros modelos de aprendizaje automático den lugar a sistemas de predicción más precisos.

Palabras Clave

Calidad del Aire Madrid, Tráfico Madrid, Predicciones metereológicas, NOAA, GFS, Aprendizaje Automático, Regresión

Abstract

Pollution is an increasing thread for human health and a environmental concern for the future. In particular, the nitrogen dioxide, which is generated mainly by traffic in urban areas, is reaching levels that are dangerous for health. Governments of diverse countries and cities have already established laws in order to reduce these pollution levels. The city of Madrid, which has to to obey the European laws that establish maximum levels of NO_2 in cities, is a remarkable case. Therefore, accurate predictions of NO_2 concentration are interesting in order to improve management asociated with the anti-pollution measures, which include speed limits in some roads or even access and parking restrictions in the center of Madrid. To achieve this goal, machine learning plays a crucial role. Machine learning is a branch of artificial intelligence that explores the study and construction of predictive modelling, which are essential in order to know in advance the levels of NO_2.

The objective of this work is to create a predictive model for NO_2 concentration. For this purpose, historic NO_2 concentration data, real traffic data and metereology prediction data in the city of Madrid are used in this work. Before generating the model, the characteristics of the data contained in are carefully analyzed. The local government of Madrid offers for free the information about traffic and air quality, which are registered in numerous sensors and stations placed in the city. The metereological predictions are obtained freely from the National Ocenaic and Atmosferic Administration (NOAA) of the United States.

In this first work the model is restricted to four air quality stations placed in Plaza de España, Escuelas Aguirre, Méndez Álvaro and Plaza del Carmen, using the traffic load from their nearest traffic sensors. Metereological features of temperature, wind speed, pressure and relative humidity are used as well. Based on this information, it is possible to predict the NO_2 concentration with relative accuracy. Using a linear regression algorithm with the ElasticNet regularization a mean relative error (MRE or MAPE) of 20 % is obtained. Additionally it has been shown that models that include features related to traffic are better than those that only use historic NO_2 concentration data. Furthermore, an alert classification model has been implemented. This model, uses the XGBoost algorithm, and predicts accurately the majority of alerts.

The results obtained in this TFG are promising, and it is expected that the inclusion of new features or the use of different models will give as a result more accurate prediction systems.

Key words

Madrid Air Quality, Madrid Traffic, Metereological predictions, NOAA, GFS, Machine Learning, Regression

Agradecimientos

En primer lugar, quiero dar las gracias a mi tutora Irene por su entusiasmo constante, sus ánimos y los esfuerzos que ha hecho para cuidar cada detalle del trabajo y hacerlo a tiempo. Sin ella este trabajo no habría sido posible.

En segundo lugar agradezco a Sara haber sido la mejor compañera de prácticas que se podía tener, gracias haberme escuchado y haberme ayudado tanto cuando me hacía falta. Especialmente gracias por ayudarme con los detalles de Látex.

Gracias también a mis compañeros de carrera, que, aunque ahora nos veamos poco, siempre nos hemos apoyado. También agradezco a mis padres y a mi hermano el apoyo que me han dado, no sólo ahora si no durante toda la carrera.

Por supuesto, muchas gracias a Alberto Torres que me dejó utilizar sus scripts para descargar los datos de las predicciones metereológicas, sin su ayuda habría tardado dos meses más en acabar este trabajo.

Me gustaría agradecer también a Carlos Santa Cruz la oportunidad de trabajar en el IIC junto a él, donde he aprendido muchas de las cosas necesarias para realizar este trabajo. Por último, le doy las gracias a José Dorronsoro por ofrecerse a ser mi ponente y por su trato siempre atento.

Índice general

Índice de Figuras

Índice de Tablas

1 Introducción

1.1. Motivación del proyecto

En los últimos años el problema de la contaminación ha adquirido una gran relevancia. La contaminación del aire afecta principalmente a áreas urbanas y en algunos países ha provocado una situación crítica [4, 5]. Los principales contaminantes que afectan a las ciudades son el ozono (O_3), el dióxido de carbono (CO_2), las partículas en suspensión (PM_{10} o $PM_{2.5}$), el dióxido de azufre (SO_2) y los óxidos de nitrógeno (NO y NO_2). Diversos artículos [6, 7, 8], discuten la posibilidad de que la exposición puntual a la contaminación del aire pueda causar problemas de salud como irritación de los ojos, dificultades respiratorias o problemas cardiovasculares, mientras que la exposición prolongada afecta al sistema neurologico, inmune, respiratorio y reproductor, pudiendo causar cáncer o incluso la muerte prematura. El dióxido de nitrógeno, en concreto, es especialmente nocivo.

Con el objetivo de paliar los efectos del NO_2 los gobiernos de todo el mundo han adoptado medidas que, entre otras cosas, tratan de reducir la contaminación del aire en las ciudades, en particular la concentración de NO_2. La normativa de la Unión Europea obliga a los países que la forman a mantener un nivel medio de NO_2 por debajo de los $40\mu g/m^3$ y no superar más de 18 horas al año los 200 $\mu g/m^3$. Madrid es una de las pocas ciudades europeas que no cumple esta normativa, por ello el ayuntamiento ha creado un protocolo de alertas de alta contaminación [9]. En este protocolo, cuando el nivel de concentración de $N0_2$ alcanza unos límites determinados se restringe el tráfico hasta que se vuelve a niveles de concentración permitidos.

Por todo ello, sería deseable disponer de un sistema que predijera los niveles de concentración de $N0_2$ con una cierta antelación, dando un mayor margen a la población y entidades públicas para prepararse ante restricciones del tráfico y también permitiendo tomar medidas contra la contaminación de manera preventiva. Para la predicción de niveles de contaminantes es necesario recurrir al Aprendizaje Automático, un campo de la Inteligencia Artificial que trata de extraer patrones y relaciones de un conjunto de datos sin necesidad un conocimiento experto sobre ellos. Con la información y los datos adecuados, aplicando técnicas de aprendizaje automático, se podría desarrollar el modelo predictor de la concentración del $N0_2$. Un modelo deber teber en cuenta que en la generación de $N0_2$ en una ciudad intervienen dos factores principalmente [10]:

- El tráfico

- La metereología

El Ayuntamiento de Madrid ofrece de forma abierta los datos del tráfico y de la calidad del aire en la ciudad [11], tanto datos históricos como en tiempo real. Por otra parte, existen multitud de fuentes que ofrecen predicciones metereológicas tales como la AEMET, ECMWF o NOAA. En este caso se ha elegido la última debido a ofrece sus predicciones de forma gratuita. Este trabajo se centra un utilizar estos datos combinados con las técnicas de aprendizaje automático para predecir la concentración de NO_2.

1.2. Objetivos y enfoque

El objetivo de este trabajo es desarrollar este modelo predictor de la concentración de $N0_2$ en Madrid utilizando datos históricos de la concentración de NO_2 , datos de tráfico y predicciones metereológicas, así como analizar que variables intervienen en la formación de este contaminante. Se pretende proporcionar un modelo de regresión que prediga la concentración de contaminate en $\mu g/m^3$ y un modelo de clasificación que divida los ejemplos entre ejemplos normales y alertas, definidos en función de los umbrales de $<180\mu g/m^3$ y $>180\mu g/m^3$ respectivamente. El objetivo de la regresión es predecir la concentración horaria de NO_2 con un horizonte de predicción de 24 horas. Para el desarrollo, tanto de la parte de transformación de datos como la parte de aprendizaje automático, se utiliza Python y muchas de sus librerías entre las que cabe destacar 'scikit-learn' que implementa un amplio espectro de algoritmos de aprendizaje automático. Para la consecución del objetivo de este TFG, en primer lugar hay que conseguir los datos, transformarlos y combinarlos para construir el conjunto de aprendizaje. Una vez construido este se analizará y se construirán modelos utilizando técnicas estadísticas y de aprendizaje automático.

1.3. Metodología y plan de trabajo

En este documento se expone el trabajo realizado para el desarrollo de un sistema predictor de la concentración de NO_2. El segundo capítulo muestra una revisión del estado del arte en la predicción de contaminantes así como una descripción de las fuentes de datos y las técnicas de aprendizaje automático utilizadas. El tercer capítulo explica el proceso de desarrollo del modelo, desde la extracción de datos en las fuentes hasta la construcción de un modelo que permita extraer conclusiones de los datos. El cuarto capítulo expone el análisis y resultados obtenidos del modelo construido y de los datos utilizados. Por último, el capítulo 5 muestra las conclusiones que se pueden extraer de los resultados mostrados en el capítulo 4.

2

Predicción de la contaminación. Estado del arte

Este primer capítulo introducirá diversos aspectos relacionados con la predicción de NO_2. En la sección 2.1 se hará un repaso del estado del arte de predicción de contaminantes. A continuacíón, en la sección 2.2 se presentarán las características del contaminante que queremos predecir. Las secciones 2.3, 2.4 y 2.5 describen respectivamente las fuentes de datos utilizadas para la información del tráfico, calidad del aire y metereología. Por último, la atención se centra en los fundamentos del aprendizaje automático y en las secciones 2.6 y 2.7 se explican los métodos de clasificación y regresión utilizados en este trabajo.

Nuestro objetivo es predecir la concentración de $N0_2$ en Madrid para poder hacer predicciones horarias. Aunque es importante remarcar que, a pesar de que el modelo desarrollado se focaliza en el $N0_2$, la sección 2.1 hace un repaso de trabajos centrados en otros contaminantes [12, 13] . La razón por la que este trabajo se centra en el NO_2 es la repercusión que tiene éste en la provincia de Madrid, dada la necesitad de asegurar que los niveles del mismo son los establecidos por las leyes europeas [4, 14].

En la descripción de las fuentes de datos, este capítulo se centrará en los contenidos y organización de los datos, su resolución espacial y temporal, su tasa de actualización y formato de representación utilizado.

Finalmente se hará una introducción al aprendizaje automático, en concreto del problema de aprendizaje supervisado. También se explicarán algunos métodos de clasificación, centrándose especialmente en árboles de decisión y conjuntos de clasificadores como Random Forest y XGB, dado que han sido los modelos utilizados en este TFG.

2.1. Revisión de trabajos anteriores

Existen en la literatura artículos orientados a la predicción de contaminantes en diferentes ciudades como Pekín, Londres, Oporto, Szeged, Lucknow o Kuwait. Estos trabajos son diversos en cuanto al contaminante a predecir, las variables explicativas a utilizar y el método de aprendizaje automático adoptado. Algunos trabajos se centran en la predicción de $N0_2$ como el de Szeged [15] o el de Londres [5]. Las variables utilizadas han sido diversas, utilizándose en todos los casos variables metereológicas como la humedad relativa o la velocidad del viento. También

es habitual que se use alguna medida de la densidad de tráfico en la zona o valores pasados del contaminante que se quiere predecir. También se utilizan distintos métodos de clasificación como máquinas de vectores soporte, árboles de decisión, redes neuronales o regresión logística. En el caso de la regresión se usan métodos como la regresión lineal.

Toda la información concerniente a los distintos artículos que se han revisado se encuentra en la Tabla 2.1.

2.2. Características del $N0_2$

El tráfico es una de las princiaples causas en la formación del NO_2, este se forma a partir del NO u óxido nitroso, este es un gas inoloro e incoloro que se produce en la combustión de combustibles, por ejemplo en los coches. Una vez que el NO se mezcla con el aire, se combina rápidamente con el oxígeno formando NO_2. Como el dióxido de nitrógeno es un contaminante relacionado con el tráfico, las emisiones son más elevadas en áreas urbanas [5] .

En cuanto a los efectos del NO_2 en la salud, los estudios [18] realizados sobre poblaciones humanas indican que la exposición a largo plazo al NO_2, a los niveles que actualmente se registran en Europa, puede provocar una disminución de la función pulmonar y aumentar el riesgo de aparición de síntomas respiratorios como bronquitis aguda, tos y flema, especialmente en los niños y asmáticos.

Tambien demuestran que la exposición al NO_2 aumenta la respuesta alérgica al polen inhalado, lo que provoca graves dificultades respiratorias especialmente en personas asmáticas y con alergia.

Además, no hay pruebas respecto a la existencia de un umbral de exposición al NO_2 por debajo del cual no sea previsible ningún efecto sobre la salud.

2.3. Portal Datos Abiertos: Tráfico

Referencia	**Contaminante a predecir**	**Variables**	**Horizonte**	**Métodos**	**Lugar**	**Error**
[13]	PM2.5	Temperatura, Velocidad del viento, Humedad Relativa, Índice de tráfico (de 0 a 10), Contaminación del día anterior (alta o baja)	1 día	Logistic Regression	Pekín	9.09 %
[16]	Predicen el AQI* y CAQI**	SO2, NO2, SPM, RSPM, Temperatura, Humedad relativa, Velocidad del viento, Evaporación, Periodo de luz solar	No se especifica	SDT, DTF y DTB.	Lucknow (India)	5.62 %
[5]	NO2	Concentración media de NO2 en la última hora, concentración media de NO2 en el último día, Volumen de tráfico, Velocidad del viento, Dirección del viento	1 h	(ANN), en concreto MLP y ARIMAX. Método conjunto de ambas.	Londres	16.53 % (MAPE o MRE)
[12]	O3	CH4, NMHC, CO, CO2, NO, NO2, SO 2, Velocidad del viento, Temperatura, Humedad relativa, Radiación solar	1 h	regresión múltiple con PCA y NN	Kuwait	??
[15]	NO y NO2	NO, NO2, Temperatura, Humedad relativa, Velocidad del viento	1 día	MLP y SVR. Reducen el numero de variables con PCA.	Szeged (Hungría)	15.2 % (MAPE o MRE)
[17]	O3	O3, NO, NO2, PM10, SO2, CO, Temperatura, Humedad relativa, Velocidad del viento	1 día	Regresión múltiple y redes neuronales prealimentadas.	Oporto	??

Tabla 2.1: Tabla de revisión de trabajos anteriores sobre predicción de contaminantes.

2.3.1. Descripción de la fuente

El Portal de Datos Abiertos [11] es una iniciativa del Ayuntamiento de Madrid que está dedicada a promover el acceso a los datos del gobierno municipal e impulsar el desarrollo de herramientas creativas para atraer y servir a la ciudadanía de Madrid. En concreto se puede consultar información tanto histórica como en tiempo real del tráfico en la ciudad. Este trabajo se ha centrado únicamente en los datos históricos.

La sensorización del tráfico se efectúa por medio de diversos equipamientos que permiten la realización del conteo de vehículos junto con la obtención del grado de ocupación. Estos sistemas de detección son en su mayoría lazos electromagnéticos que se embeben en el pavimento y detectan de forma directa la masa metálica de los vehículos que pasan sobre ellos, siendo sistemas de gran calidad y precisión, si bien tienen las limitaciones que se circunscriben a la toma de datos en un único punto y además de no disponer de visión de la zona para verificar los datos que suministra. Por ello hoy en día se tiende al uso de detectores basados en sistema de visión que si bien su calidad y fiabilidad no es tan elevada sí que permiten mayores capacidades de configuración con el establecimiento de zonas de detección en vez de puntos.

La infraestructura disponible en la ciudad de Madrid consta de 7.360 detectores de vehículos: 71 de los cuales disponen de dispositivos de lectura de matrículas, 158 de ellos cuentan con sistemas ópticos de visión artificial con control desde el Centro de Gestión de Movilidad, 1.245 son específicos de vías rápidas y acceso a la ciudad y el resto de los 5.886 sistemas básicos de control de semáforos [1].

2.3.2. Contenidos

Los datos ofrecen varias medidas distintas pertinentes al tráfico, como son la intensidad, ocupación y carga de la vía, así como velocidad media de los coches de la vía. Además se proporcionan datos que son útiles para identificar el lugar y el momento de la medición como son la fecha y hora de la medición y dos identificadores unívocos de la estación que tomó la medición, $idelem$ e $identif$. También está incluido un parámetro de error, que indica si la medición ha sido correcta o ha ocurrido algún fallo en el proceso y por lo tanto se anula su validez. La descripción de los campos que se ofrecen en los datos recogidos por los sensores se describen en la Tabla 2.2.

Nombre	Tipo	Descripción
idelem	Entero	Identificación única del Punto de Medida en los sistemas de control del tráfico del Ayuntamiento de Madrid.
fecha	Fecha	Fecha y hora oficiales de Madrid con formato yyyy-mm-dd hh:mm:ss
identif	Texto	Identificador del Punto de Medida en los Sistemas de Tráfico (se proporciona por compatibilidad hacia atrás).
tipo_elem	Texto	Nombre del Tipo de Punto de Medida: Urbano o M30.
intensidad	Entero	Intensidad del Punto de Medida en el periodo de 15 minutos (vehículos/hora).
ocupacion	Entero	Tiempo de Ocupación del Punto de Medida en el periodo de 15 minutos (%).
carga	Entero	Carga de vehículos en el periodo de 15 minutos. Parámetro que tiene en cuenta intensidad, ocupación y capacidad de la vía y establece el grado de uso de la vía de 0 a 100.
vmed	Entero	Velocidad media de los vehículos en el periodo de 15 minutos (Km./h). Sólo para puntos de medida interurbanos M30.
error	Texto	Indicación de si ha habido al menos una muestra errónea o sustituida en el periodo de 15 minutos.
periodo_integracion	Entero	Número de muestras recibidas y consideradas para el periodo de integración.

Tabla 2.2: Campos de los datos de la información del tráfico utilizados, que se extraen del Portal Abierto de Datos. [2].

2.3.3. Organización de los datos y formato utilizado

Los datos se organizan en ficheros, representando cada fichero un mes de información, y en el que cada línea del fichero corresponde a una medición de un sensor y contiene todos los campos descritos en la Tabla 2.2. En un único fichero se encuentran todas las mediciones de todas las estaciones en un mes determinado. En cuanto al formato, los ficheros son texto en plano con extensión 'csv'.

2.3.4. Resolución espacial y temporal

Los diversos sistemas de control de tráfico de la ciudad de Madrid proporcionan periódicamente y de forma automática datos de todos los detectores de vehículos de los puntos de medida que controlan. La base de datos los registra e integra sobre periodos de 15 minutos.

Si el sensor no proporciona información de una de las muestras del periodo, no se registra esa información; no obstante, si el sensor proporciona información pero los parámetros de calidad de la misma no son óptimos la información se integra pero se reporta como posible error. El error puede deberse a que el sensor detecta parámetros fuera de los rangos establecidos o que alguno de los sensores que componen el punto de medida no esté operativo (por ejemplo, en un punto de medida de 4 carriles uno de los carriles no está funcionando).

Aunque la resolución temporal es de 15 minutos, tras un primer análisis de los datos se detectan periodos de tiempo, de varias horas a veces, en las que las estaciones de medición no registran datos. Este problema de continuidad en la información se tratará en el capítulo 3.

En cuanto a la resolución espacial,ésta es del orden de decenas de metros. Hay sensores colocados en la mayoría de vías principales de la ciudad de Madrid. Un ejemplo de esta resolución puede observarse en la Figura 2.1.

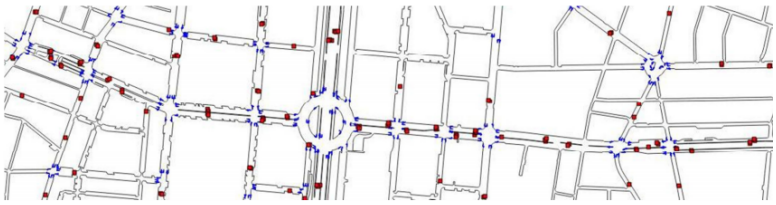


Figura 2.1: Distribución de sensorización en el eje formado por la calles Alberto Alcocer y Sor Angela de la Cruz con 86 detectores y 3 cámaras de control de tráfico del entorno [1].

2.3.5. Tasa de actualización

La tasa de actualización de los datos es de un mes. Al principio de cada mes se sube el fichero del mes anterior.

2.4. Portal Datos Abiertos: Calidad del Aire

2.4.1. Descripción de la fuente

Para los datos de calidad del aire también se usa el Portal de Datos Abiertos. En este caso también se ofrece la posibilidad de acceder a la información en tiempo real o a datos históricos, siendo estos últimos los utilizados en este trabajo.

Para llevar el control de la calidad del aire en Madrid, se utiliza un Sistema de Vigilancia, formado por 24 Estaciones Remotas automáticas que recogen la información básica para la vigilancia atmosférica. Poseen los analizadores necesarios para la medida correcta de los niveles de gases y de partículas. Los analizadores y sensores son los siguientes:

- Analizador de dióxido de azufre.
- Analizador de monóxido de carbono.
- Analizador de ozono.
- Analizador de óxidos de nitrógeno.
- Analizador de partículas en suspensión (PM10).
- Analizador de benzeno, tolueno y xileno.
- Sensor de radiación ultravioleta.

Debido a las nuevas medidas adoptadas por el ayuntamiento de Madrid [9] para reducir la concentración de NO_2 en la ciudad a los niveles que exige la legislación europea [14], este trabnajo se centra exclusivamente en el dióxido de nitrógeno a pesar de que en las estaciones se miden diversos contaminantes.

2.4.2. Contenidos

La información ofrecida contiene múltiples registros, cada uno correspondiente a una medición. En cada uno de los registros se incluyen los siguientes campos:

- **Código de estación.** Se trata de un identificador único de la estación remota.
- **Código de parámetros.** Es el campo que permite identificar el contaminante que se ha medido en el registro. Solo nos interesan los registros con mediciones de concentración de NO_2.
- **Código de técnica analítica.** Indica que técnica se ha usado para medir la concentración del contaminante. Se utiliza una única técnica para medir el NO_2, que es la quimioluminiscencia.
- **Código del periodo de análisis.** Indica si son datos horarios o diarios. En este trabajo se han usado datos horarios ya que el objetivo es predecir la concentración de NO_2 horaria con un horizonte de predicción de 24 horas.
- **Fecha.** Se incluye el año, mes y día.
- **Datos.** Los valores de concentración del contaminante se organizan en 24 tuplas, cada una correspondiente a una hora del día. Cada tupla contiene dos campos Valor y Validación, con la siguiente información:
  - **Valor.** Valor de concentración del contaminante en las unidades indicadas. En el caso del NO_2 son $\mu g/m^3$.
  - **Validación.** Se trata de un código alfanumérico que solo toma dos valores: 'V' si la medición es correcta y 'N' si no lo es.

2.4.3. Organización de los datos y formato utilizado

Todos los datos se organizan en ficheros de texto plano, sin embargo hay dos resoluciones, datos diarios o datos horarios. Como el desarrollo de este TFG se ha hecho sobre los datos horarios se expone el formato de estos. Cada fichero representa un mes y cada línea del fichero corresponde a un registro. En cada registro del fichero están indicados el día y todas las mediciones de cada hora correspondiente a ese día, además no se usan separadores para distinguir distintos campos si no que se usan campos de longitud fija. Para mayor claridad se representa en la Tabla 2.3 el formato de cada registro.

Nombre Campo	**Tipo**	**Formato**
Código de estación	Numérico	8 dígitos
Código de parámetros	Numérico	2 dígitos
Código técnica analítica	Numérico	2 dígitos
Código periodo análisis	Numérico	2 dígitos
Año	Numérico	2 dígitos
Mes	Numérico	2 dígitos
Día	Numérico	2 dígitos
Hora 1 - Valor	Numérico	5 dígitos
Hora 1 - Validación	Alfanumérico	1 dígito
...	...	...
...	...	...
Hora 24 - Valor	Numérico	5 dígitos
Hora 24 - Validación	Alfanumérico	5 dígitos

Tabla 2.3: Campos de los datos sobre la calidad del aire que se ofrecen en el Portal de Datos del Ayuntamiento de Madrid. Todas las mediciones de un mismo día se escriben en distintos campos y en la misma línea [3].

2.4.4. Resolución espacial y temporal

La resolución espacial de los datos de calidad del aire es del orden de kilómetros, con 24 estaciones de medición repartidas por Madrid. La resolución temporal es también la que necesitamos, ya que el objetivo es predecir la concentración de NO_2 cada hora.

2.4.5. Tasa de actualización

La tasa de actualización de los datos es de un mes. Cada mes se publica el fichero de registros correspondientes al mes anterior. Además, también se ofrecen datos en tiempo real que serían muy útiles si quisiera implementarse un sistema de predicción en tiempo real.

2.5. Global Forecast System del NOAA (GFS)

2.5.1. Descripción de la fuente

La Administración Nacional Oceánica y Atmosférica (National Oceanic and Atmospheric Administration, NOAA) es una agencia científica del Departamento de Comercio de los Estados Unidos que tiene varias funciones:

- Proporciona productos de información ambiental relativos al estado de los oceános y la atmósfera. En particular proporciona alertas y pronósticos como por ejemplo el *Global Forecast System* (GFS), que es el que se ha utilizado en este caso.
- Provee servicios de información ambiental.
- Es fuente de información de investigaciones científicas aplicadas.

El Sistema de Predicción Global (GFS) es un modelo de predicción del tiempo atmosférico producido por los Centros Nacionales de Predicciones Ambientales (*National Centers for Environmental Prediction*, NCEP).

Nombre	**Descripción**
Timestamp	Fecha y hora de la que se hace la predicción metereológica.
Relative_humidity_0-*i*	Humedad relativa a nivel de la superficie.
U_component_of_wind_0-*i*	Componente del viento paralela al ecuador a nivel de la superficie.
V_component_of_wind_0-*i*	Componente del viento paralela al ecuador a nivel de la superficie.
Temperature_0-*i*	Temperatura al nivel de la superficie.
Pressure_surface-*i*	Presión a nivel de la superficie.

Tabla 2.4: Campos de los datos de predicciones metereológicas

2.5.2. Contenidos

Los contenidos del GFS son muy amplios y diversos. Docenas de variables atmosféricas y a nivel del mar están disponibles en este conjunto de datos, desde temperaturas, dirección y velocidad del viento, precipitaciones hasta concentración de ozono atmosférico.

En nuestro caso se ha seleccionado un subconjunto de variables que se consideran útiles para el problema que se pretende resolver basándose en los artículos estudiados en la Sección 2.1. Estas variables son:

- Temperatura (K).
- Presión (Pa).
- Humedad relativa (%).
- Componentes del viento (m/s).

Para reducir la complejidad del problema, todas estas variables se han tomado a nivel de suelo. La inclusión de nuevas variables se deja como trabajo futuro. Los campos concretos que se utilizan en el desarrollo de este TFG se muestran en la Tabla 2.4. Como hay distintos puntos geográficos en los que se realiza la predicción, se han utilizado los puntos más cercanos a Madrid, cuya unicación pueden consultarse en la figura 2.2. Por tanto se dispone de dos valores para cada una de las variables metereológicas del modelo. Estas variables se distinguen añadiendo los sufijos 0 y 1 a su nombre según la medición se corresponda con el punto situado al oeste o al este de Madrid, respectivamente.

2.5.3. Resolución espacial y temporal

La resolución temporal de los datos del NOAA es de 3 horas. Cada día se calculan 4 modelos predictivos globales: a las 00:00, a las 06:00, a las 12:00 y a las 18:00, todos los tiempos medidos en horario UTC . Estos modelos proporcionan predicciones metereológicas, de múltiples variables, en intervalos de tres horas a partir de la hora a la que se publica el modelo predictivo. Por ejemplo, el modelo de las 06:00 tendrá la predicción de todas las variables metereológicas a las 09:00, a las 12:00, a las 15:00 y así hasta 40 intervalos de 3 horas. En nuestro desarrollo se ha tenido en cuenta el modelo de las 00:00 y los valores de predicciones de las primeras 24 horas. Es decir, de cada día tenemos las predicciones de las variables metereológicas a las 03:00, a las 06:00, a las 09:00, a las 12:00, a las 15:00, a las 18:00, a las 21:00 y a las 24:00. Se utiliza el modelo de las 00:00 porque simplifica las operaciones temporales, y sólo se utiliza la previsión de las siguientes 24 horas porque así podemos asegurar un cierto grado de confianza en la predicción.

En cuanto a la resolución espacial, para generar el GFS se divide el mapa terrestre en una rejilla y se calcula el modelo en cada punto de la rejilla. La resolución de esta rejilla es de 5° [19].

Esto provoca que aunque la distancia entre dos puntos a la misma latitud y distinta longitud es constante, la distancia entre dos puntos de la misma longitud y distinta latitud es variable, siendo mucho mayor en el ecuador que en los polos.

2.5.4. Organización de los datos y formato

Como se explica en la sección anterior, cada día se generan 4 modelos predictivos que se guardan en un fichero cada uno . Se puede acceder a todos los modelos generados desde el 27/5/2015. El formato que se utiliza para almacenar los datos es GRIB. GRIB es un estándar usado comúnmente en metereología tanto para almacenar predicciones como datos históricos. Los ficheros de GRIB son una colección de datos autocontenidos, es decir sin referencias a otros ficheros, en 2 dimensiones. Cada registro consta de dos partes, la cabecera, que describe las características del registro, y la información en sí en formato binario. En el caso concreto del modelo predictivo que se usa, cada predicción en un determinado momento de una cierta variable se almacena como una matriz de 720×361 donde la fila indica la longitud y la columna indica la latitud. De esta forma el elemento que aparece en la primera fila y primera columna representa el punto (0E, 90N). Las coordenadas que se han elegido para Madrid son (40.5N, 4W) y (40.5N, 3.5W) ya que son los dos puntos de la rejilla más cercanos a la ciudad.

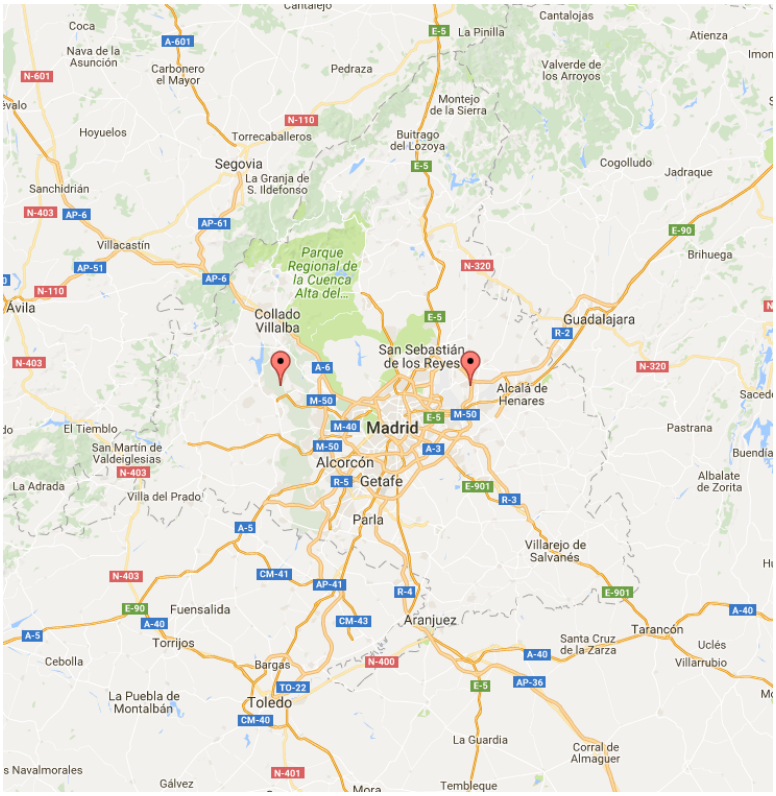


Figura 2.2: Mapa de Madrid en el que se marcan las dos coordenadas en las que se calcula la predicción de metereología que luego se usará para el modelo.

2.5.5. Tasa de actualización y detalles de acceso

La tasa de actualización es de 6 horas ya que cada día se calculan cuatro nuevos modelos de predicción. Se puede acceder a estos modelos mediante una petición http al servidor y se pueden

descargar los datos en formato ascii o grib. En este TFG se ha elegido la última opción ya que los datos ocupan un espacio menor y es más rapido el acceso a datos nuevos.

2.6. Algoritmos de Clasificación Supervisada

El aprendizaje automático es una rama de la inteligencia artificial cuyo objetivo es el de aprender y descubrir las reglas y estructuras subyacentes a los datos. En nuestro caso nos centraremos en los problemas de clasificación supervisada, partiendo de n ejemplos o muestras definidos como $\vec{x_i}$ con unas etiquetas y_i para $i = 1, ..., n$. A partir de estos datos, en la fase de entrenamiento se genera un clasificador cuyo objetivo es inferir una función f que permita predecir la etiqueta $\hat{y_j}$, de un nuevo elemento $\vec{x_j}$ no visto en la fase de entrenamiento. A su vez, cada ejemplo o muestra está formado por d variables $\{x_i^1, ..., x_i^d\}$. Por tanto, un clasificador es un algoritmo que *aprende* patrones a partir de un conjunto de datos de entrenamiento, sin conocimiento previo del problema, y cuyo resultado es una función:

$$f : \mathbb{R}^d \rightarrow \{y_1, ..., y_l\}$$

La función f asigna a cada patrón $\vec{x} \in \mathbb{R}^d$ una etiqueta $y_k \in \{y_1, ..., y_l\}$ intentando cometer el menor error posible. Este error se mide típicamente como el cociente entre el número de ejemplos predichos incorrectamente y el número total de ejemplos. Para clarificar posteriores explicaciones es necesario notar que los datos que se usan para entrenar el clasificador, y también para predecir nuevas etiquetas, se guardan en forma de matriz. En esta matriz, cada fila representa un elemento (muestra o patrón) y cada columna representa una variable, de manera que todos los patrones tienen las mismas variables aunque sean con valores distintos. Para este trabajo se han usado algoritmos basados en conjuntos de clasificadores como son Random Forest [20] o métodos de boosting como XGB (extreme gradient boosting) [21]. Ambos algoritmos entrenan conjuntos de clasificadores individuales, cada uno de los cuales uno predice una etiqueta para un mismo patrón. La predicción final del algoritmo es el voto por mayoría en el conjunto de clasificadores. En la subsección 2.6.2 se ampliará la explicación de estos métodos.

2.6.1. Árboles de decisión

En los conjuntos de clasificadores considerados se utiliza como clasificador base los árboles de decisión, por ello se comenzará dando una breve explicación de este método de aprendizaje automático. Hay que remarcar que aunque los árboles de decisión pueden usarse con variables categóricas, como los datos con los que se construye el modelo predictivo son reales, nos centraremos en aquellos que se usan para variables reales. Un árbol de decisión T es una secuencia ordenada de preguntas sobre las variables que definen cada patrón $\vec{x_i}$ con el objetivo de asignarle una determinada clase y_j. En un árbol de decisión, a los nodos internos se les asigna una pregunta o regla que divide el conjunto de patrones en dos subconjuntos. De esta manera, el conjunto inicial $\mathcal{L}$ se va subdividiendo hasta alcanzar los nodos hoja, que constituyen una partición disjunta del espacio de características inicial, ya que cada ejemplo puede seguir un único camino al recorrer el árbol desde la raíz hasta las hojas. Las preguntas que se asignan a cada nodo interior son del tipo ¿$x_i^k \leq \alpha$?, donde x_i^k es el valor de la variable k-ésima del ejemplo $\vec{x_i}$ a clasificar y recibe el nombre de variable de corte, además alpha es un número real y recibe el nombre de valor de corte. Para poder implementar el algoritmo es necesario definir tres cosas:

1. La función de impureza que se utiliza para medir la bondad de las variables y valores de corte, asi como algoritmo y criterio que se aplica para encontrar los óptimos.

2. Un criterio para asignar una clase a cada nodo hoja.
3. Un criterio para saber en cada nodo si seguir expandiendo el árbol o marcar el nodo como hoja (Criterio de parada).

A continuación se detalla como el algoritmo CART [22] resuelve cada uno de estos problemas.

Función de impureza. Selección de variable y valor de corte en cada nodo

El objetivo principal del árbol de decisión es que cada hijo derecho e izquierdo sean lo más puros posibles respecto a su nodo padre, en cuanto a la separación en las clases. Es decir, un nodo es más puro cuanto más separación haya de las clases en dicho nodo, y se pretende que los descendientes separen mejor las clases que sus predecesores. Sea un problema de clasificación con J clases distintas, entonces la función de impureza Φ definida sobre $p_1, ..., p_J$, la proporción de las clases en el nodo, debe cumplir las siguientes tres propiedades:

1. Que alcance su máximo en $(1/J, ..., 1/J)$, vector de J componentes
2. Que alcance su mínimo en los vectores con una coordenada con valor 1 y las J-1 coordenadas restantes con valor 0.
3. Que sea simétrica para todas las clases.

Con estas propiedades existen diversas funciones. Definiendo la impureza en el nodo i como $i(t) = \Phi(p(1|t), ..., p(J|t)) = \Phi(p_1, ..., p_J)$ podemos destacar las siguientes medidas.

1. GINI
$$\Phi(p_1, ..., p_J) = \sum_{i=1}^{J} p_i(1 - pi)$$
2. Entropía
$$\Phi(p_1, ..., p_J) = -\sum_{i=1}^{J} p_i \log_2(p_i)$$

La medida utilizada en CART y la que se ha elegido para este TFG es GINI.

Una vez que hemos definido la función de impureza, se elegirá aquella variable y valor de corte que maximicen la diferencia entre la impureza del padre y de los hijos. Es decir, el corte óptimo será aquel que maximice:

$$\Delta i(t) = i(t) - p_L \cdot i(t_L) - p_R i(t_R)$$

Donde $i(t_L), i(t_R)$ son las impurezas del hijo izquierdo y derecho y p_L, p_R son la proporción de datos que va hacia cada uno de los dos hijos.

Es importante conocer también cómo se construye un árbol de decisión cuando los ejemplos tienen distintos pesos. Supongamos que en un nodo cada ejemplo $\vec{x_i}$ del conjunto de entrenamiento tiene asignado un peso w_i y una etiqueta y_i para $i = 1, ..., n$. Además, definimos la función *class* como la que le asigna a cada elemento del conjunto de entrenamiento $\vec{x_i}$ su clase y_i entonces, en vez de definir $p_1, ..., p_J$ como la proporción de las clases en el nodo, lo hacemos de la siguiente manera:

$$p_k = \frac{\sum_{i=1}^{n} w_i[class(x_i) == y_k]}{\sum_{i=1}^{n} w_i},$$

donde $[class(x_i) == y_k]$ indica la función indicatriz de la clase y_k.

Asignación de clase a los nodos hoja

Una vez que un nodo ha sido elegido como nodo hoja la asignación de la clase es inmediata: se le asignará aquella clase en mayor proporción en el nodo. Es decir:

$$t \text{ es etiquetado como } y_k \iff y_k = \underset{\{y_1,\dots,y_J\}}{\arg\max}\, p(y_i|t)$$

Criterio de parada

Hay diversos criterios de parada, dentro de los cuáles existen dos muy simples. El primero es definir una profundidad máxima del árbol. El segundo es desarrollar el árbol hasta que todos los nodos tengan impureza 0, o hasta que no se pueda hacer menor la medida GINI. El problema de este segundo método es que los árboles de decisión pierden capacidad de generalización. Existen otros criterios de parada más complejos que no se tendrán en cuenta en este trabajo, así como criterios de poda para reducir el riesgo de sobreajuste.

2.6.2. Conjuntos de clasificadores

Los conjuntos de clasificadores son sistemas en los que varios clasificadores individuales, que se denominan clasificadores base, son combinados para predecir las etiquetas de nuevos ejemplos, con el objetivo de que el error cometido por el conjunto sea menor que el de cada clasificador individualmente. En general, el rendimiento obtenido por un conjunto de clasificadores será mayor que el obtenido por cada uno de sus elementos bajo las siguientes condiciones [23]:

- **Precisión.** El error cometido por cada uno de los clasificadores tiene que ser menor que el que se obtendría si se clasificara aleatoriamente.
- **Diversidad.** Los errores cometidos entre los clasificadores individuales no deben estar correlacionados. La diversidad es fundamental ya que si todos los elementos del conjunto son prácticamente iguales, el resultado conjunto no distaría mucho del resultado de cada elemento.

Aunque hay diversos métodos para conseguir una mayor diversidad nos centraremos en dos principalmente:

- **Bagging** (Boostrap aggregation). Consiste en seleccionar subconjuntos del conjunto original de muestras de forma aleatoria y con reemplazamiento con tantos ejemplos como haya en la muestra original, y entrenar cada clasificador con uno de estos subconjuntos. De esta manera se consigue que algunos ejemplos no aparezcan en el conjunto de entrenamiento mientras que otros aparecen varias veces.
- **Boosting** Se trata una de las técnicas más eficaces en la construcción de conjuntos de clasificadores. Consiste en que cada clasificador se construye en base a los resultados obtenidos por los clasificares anteriores, modificando los pesos de los ejemplos de entrenamiento. Patrones que han sido clasificados erróneamente tendrán una mayor importancia en la construcción del siguiente clasificador, consiguiendo así que los clasificadores sucesivos se centren en los elementos del conjunto de entrenamiento más dificiles de clasificar.

Para combinar los resultados de los clasificadores individuales se usa una votación ponderada (boosting) o una votación no ponderada (Random Forest.)

Random Forest

El Random Forest es un modelo de Aprendizaje Automático basado en conjuntos de clasificadores, donde los clasificadores individuales son árboles de decisión. Este modelo presenta una ventaja a la hora de lograr diversidad en el conjunto de clasificadores ya que en los árboles de decisión se puede añadir aleatoriedad a la construcción de los mismos. En vez de seleccionar el mejor corte de entre todas las variables, en cada nodo se seleccionan F variables de forma aleatoria y después se elige el mejor corte dentro de este subconjunto. Uno de los algoritmos más utilizados para el entrenamiento del Random Forest es el algoritmo 1 propuesto en [20] que aúna bagging y aleatoriedad en las variables escogidas en la construcción de los árboles.

Algorithm 1: Algoritmo Forest-RI

Data: Muestra $\mathcal{L} = \{(\vec{x_n}, y_n), n = 1, ..., N,\ \vec{x_n} \in \mathbb{R}^d,\ y_n \in \{y_1, ..., y_J\}\}$
T: número de árboles a construir.
F: número de variables a considerar en cada nodo
Result: Random Forest

```
for t := 1 to T do
    L_bt = MuestreoBootstrap(L)
    c_t = ConstruyeArbolAleatorio(L_bt, F)
end
```

En el Algoritmo 1 MuestreoBootstrap es una función que devuelve una muestra de N elementos con reemplazamiento y ConstruyeArbolAleatorio es una función que construye un árbol eligiendo F variables aleatorias en cada nodo. Una vez que se han construido todos los árboles, la decisión final del algoritmo se define como la clase más votada entre los clasificadores individuales del modelo.

2.6.3. XGBoost

XGBoost [21] es un método para realizar el algoritmo de boosting sobre árboles de decisión de una manera rápida y eficiente computacionalmente. Además es uno de los métodos más utilizados por los ganadores de concursos organizados por *Kaggle*.

2.7. Algoritmos de Regresión

Los algoritmos de regresión son un subconjunto de las técnicas del aprendizaje automático en el que, en vez de asignar a cada ejemplo una etiqueta dentro de un conjunto finito de etiquetas, se debe predecir un valor continuo para dicho ejemplo. La matriz de datos se expresa igual que en los algoritmos de clasificación. Es decir, dado un conjunto de muestras $\mathcal{L}$, donde cada muestra tiene d variables. Se representa como una matriz X en el que cada fila representa un ejemplo y cada columna es una variable. Además, en el caso de la regresión ,denotamos como X_i para $i = 1, ..., d$ a la variable aleatoria asociada a la i-ésima variable de la muestra. Mientras que en la clasificación se clasifican los ejemplos de $\mathcal{L}$ dentro de un conjunto finito de grupos asignándoles una etiqueta $y_i \in \{y_1, ..., y_J\}$, en la regresión se predice un valor continuo Y. Este valor Y continuo recibe el nombre de variable explicada o dependiente, mientras que cada una de las variables de los ejemplos $X_1, ..., X_d$ se llaman variables explicativas o regresoras.

2.7.1. Regresión lineal

La regresión lineal es un modelo matemático que parte de la hipótesis de que la variable explicada puede expresarse como una combinación lineal de las variables regresoras.

$$Y_i = \beta_0 + \beta_1 x_i^1 + ...\beta_d x_i^d + \epsilon_i$$

Donde $\vec{\epsilon} = (\epsilon_1, ..., \epsilon_N) \sim N(0, \sigma^2 I)$ representa la componente no lineal del modelo. De esta definición de ϵ se puede deducir que la media de los errores es 0, es decir no se sobreestima ni se subestima sistemáticamente. Y como la varianza es $\sigma^2 I$ se observa que todos las errores tienen la misma varianza y además son linealmente independientes entre sí. Se busca entonces un modelo:

$$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} x_i^1 + ... + \hat{\beta_d} x_i^d + e_i \ ,$$

donde $e_i = Y_i - \hat{Y_i}$ Se calculan entonces los coeficientes $\hat{\beta_0}, ..., \hat{\beta_d}$ que hacen $\sum_{i=1}^{N}(e_i^2)$ mínimo. Es decir, se busca el vector $\vec{\hat{\beta}} = (\hat{\beta_1}, ..., \hat{\beta_d})]$ y el término independiente $\hat{\beta_0}$ de manera que

$$\{\hat{\beta_0}, \vec{\hat{\beta}}\} = \arg\min_{\beta_0, \vec{\beta}}(Y_i - \beta_0 - \vec{\beta}\vec{x_i}) \ .$$

Este método se denomina método de mínimos cuadrados. Sea X la matriz donde cada columna representa una variable y cada fila un ejemplo de $\mathcal{L}$. Entonces se puede demostrar que el estimador de mínimos cuadrados es:

$$\hat{\beta} = (XX^T)^{-1}X^T Y \ ,$$

donde X^T es la matriz traspuesta de X.

Regularización de Lasso

Con el objetivo de mejorar la precisión de la regresión lineal al dar mayor peso a las variables más importantes y a su vez realizar una selección de variables, se utiliza la regularización de Lasso. Se hallan $\hat{\beta_0}, \hat{\beta}$ de la siguiente manera:

$$\{\hat{\beta_0}, \vec{\hat{\beta}}\} = \arg\min_{\beta_0, \vec{\beta}}(Y_i - \beta_0 - \vec{\beta}\vec{x_i} - \lambda\|\beta\|_{L_1})$$

La idea es penalizar los vectores β de mayor modulo y, al usar la normal $L1$ se consigue que los coeficientes de las variables menos relevantes se hagan 0 [24]. El parámetro λ óptimo para cada problema suele ajustarse utilizando una búsqueda en rejilla y usando validación cruzada sobre el conjunto de muestras.

Regularización de ElasticNet

La regularización de Lasso presenta varias limitaciones. Por ejemplo, en conjuntos de datos con muchas dimensiones (d) y con pocos ejemplos (n), Lasso cogerá como mucho n coeficientes distintos de 0. Además, si un grupo de variables están muy correladas, Lasso tiende a escoger una única variable de este grupo e ignorar el resto. Para superar estas limitaciones se introduce un nuevo término que es la norma L_2 de β en la ecuación:

$$\{\hat{\beta_0}, \vec{\hat{\beta}}\} = \arg\min_{\beta_0, \vec{\beta}}(Y_i - \beta_0 - \vec{\beta}\vec{x_i} - \lambda_1\|\beta\|_{L_1} - \lambda_2\|\beta\|_{L_2})$$

Se ajusta λ igual que en Lasso.

2.7.2. Random Forest Regressor

En la regresión lineal se afronta el problema de forma global, utilizando todo el espacio de características. En problemas con muchas variables que interactúen de forma no lineal, construir un modelo lineal puede dar rendimientos pobres. Además, en problemas en las que las variables no son independientes se tienen problemas derivados de la inversión de matrices singulares. Para resolver estos problemas pueden usarse los árboles de regresión. La idea es dividir el espacio de características en particiones disjuntas para construir un modelo lineal en cada partición. Para construir el árbol y subdividir el espacio, el proceso es similar al que se sigue en los árboles de decisión, la diferencia es que ahora en vez de minimizar la impureza de los nodos, se trata de maximizar la información de la partición P con la variable dependiente Y. Sea $I(P;Y)$ esta información, el objetivo es encontrar la variable y valor de corte que maximicen la diferencia entre la información con Y en las particiones correspondientes a los nodos hijos y la correspondiente al nodo padre. Es decir:

$$\textit{Se elige como pregunta } (X_r < \alpha_s) \iff X_r, \alpha_s = \arg\max_{X_i,\alpha_j} p_L I(P_L;Y) + p_R I(P_R;Y) - I(P;Y)$$

Donde p_R y p_L son la proporción de ejemplos que se va a cada uno de los hijos y P_L y P_R son las particiones correspondientes a los hijos izquierdo y derecho respectivamente. De esta manera, para construir un árbol se comienza en el nodo raíz y se van construyendo nodos hijos y dividiendo el espacio hasta que se cumpla una cierta condición de parada, como una profundidad máxima. En los nodos hoja se calcula una regresión lineal y se le asigna ese valor al ejemplo. En el caso del *Random Forest Regressor* se hace la media de los valores que predice cada árbol individualmente.

3

Sistema, diseño y desarrollo

En este capítulo se describe el proceso al que se someten los datos utilizados desde que se accede a las fuentes de datos elegidas hasta que se combinan para formar un modelo. Los distintos datos se unen entre sí utilizando la fecha y hora, de manera que cada hora se tenga información de tráfico, calidad del aire y metereología. Antes de poder combinarlas, cada una de las fuentes de datos necesita una serie de pasos que son distintos, por ello se divide el capítulo en una sección por tipo de datos. En primer lugar se explicará el proceso llevado a cabo para los datos del tráfico (Sección 3.0.1), después el que se aplica sobre los datos de calidad del aire (Sección 3.0.2) y, por último, se describirá el tratamiento llevado a cabo sobre los datos metereológicos (Sección 3.0.3).

Como resultado de estas transformaciones sobre los datos originales se obtendrá una matriz que servirá de entrada a los modelos de aprendizaje automático. Antes de dar una visión simplificada de esta matriz es necesario introducir las siguientes definiciones:

- Sea S la estación de calidad del aire sobre la que se construye el modelo. Se va a construir un modelo individual para cada estación de calidad del aire (cuándo se utilicen distintas estaciones habrá L distintas y se nombran $S_1, ..., S_L$).
- Sean M_1 y M_2 puntos en los que se predicen las variables metereológicas (Como se muestra en la Sección 2.5, son los dos únicos puntos de predicción suficientemente cercanos a la ciudad de Madrid para ser tenidos en cuenta).
- Sean $T_1, T_2, ..., T_K$ los K puntos de medida de tráfico más cercanos a S.
- Sean h el horizonte temporal para el que se quiere predecir la concentración de NO_2, p la profundidad temporal de la que se quiere tener información, s el 'step' o paso temporal entre cada medición (podemos suponer 1h ya que se pretende hacer una predicción horaria) y t_0 el tiempo en el que se quiere predecir la concentración de NO_2.
- Sean $v_1, v_2, \ldots, v_d$ las variables metereológicas. v_i^1 y v_i^2 para $i = 1, ..., d$ son las variables predichas en los puntos M_1 y M_2 respectivamente. Además, $v_i^j(t)$ indica la variable i-ésima medida en la estación j-ésima en el tiempo t.
- Sea $c_i(t)$ para $i = 1, ..., K$ la carga del tráfico medida en el i-ésimo punto de medida del tráfico más cercano a S en el momento t.

- Sea $C(t)$ la concentración de NO_2 medida en la estación S en el momento t

Una vez definidos los conceptos más básicos hay que definir algunos más complejos:

- Sea $\hat{T}_i(t,p) = c_i(t), c_i(t-s), \ldots, c_i(t-ps)$ para $i = 1, \ldots, K$ las cargas de tráfico medidas en la estación T_i en el tiempo t y con una profundidad p, es decir además de la medición en el tiempo t se tienen medidas de p momentos anteriores.
- Sea $\hat{M}_j(t,p) = v_1^j(t), v_1^j(t-s), \ldots, v_1^j(t-ps), \ldots, v_d^j(t), v_d^j(t-s), \ldots, v_d^j(t-ps)$ para $j = 1, 2$ las d predicciones de variables metereológicas para la estación M_j en el tiempo t y con una profundidad p.

Entonces la matriz que se quiere conseguir es:

$$\left[\begin{array}{cccccc|c} C(t_0-(n+h)s) & \hat{T_1}(t_0-ns) & \ldots & \hat{T_K}(t_0-ns) & \hat{M_1}(t_0-ns) & \hat{M_2}(t_0-ns) & C(t_0-ns) \\ C(t_0-(n-1+h)s) & \hat{T_1}(t_0-(n-1)s) & \ldots & \hat{T_K}(t_0-(n-1)s) & \hat{M_1}(t_0-(n-1)s) & \hat{M_2}(t_0-(n-1)s) & C(t_0-(n-1)s) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ C(t_0-(1+h)s) & \hat{T_1}(t_0-s) & \ldots & \hat{T_K}(t_0-s) & \hat{M_1}(t_0-s) & \hat{M_2}(t_0-s) & C(t_0-s) \\ C(t_0-hs) & \hat{T_1}(t_0) & \ldots & \hat{T_K}(t_0) & \hat{M_1}(t_0) & \hat{M_2}(t_0) & C(t_0) \end{array}\right],$$

donde la última columna representa la variable objetivo a predecir, esto es, la concentración de NO_2 en un determinado instante de tiempo.

3.0.1. Datos de trafico

Los datos en crudo con información sobre el tráfico en Madrid tienen, el formato explicado en la Tabla 2.2. Con el objetivo de poder combinarlo con el resto de datos se ha seguido una serie de pasos que se explican a continuación.

Limpieza de los datos

Se han eliminado las columnas de *vmed*, *ocupacion*, *intensidad* y *periodo_integracion*. Esto se hace para simplificar el modelo ya que la *carga*, es el dato es el dato que queremos utilizar como indicador del tráfico en la vía. Esta elección se hace porque la *carga* es una medida que asigna a la carretera una puntuación de 0 a 100 teniendo en cuenta la intensidad, ocupación y capacidad de la vía También se elimina la columna de *identif* ya que con la de *idelem* es suficiente para identificar cada estación. Además se han eliminado aquellas mediciones con la columna de error distinta de 'N' ya que otro valor indicaba un error en la medición, algo que no es deseable tener para un modelo predictivo. En el siguiente apartado se explicará cómo se solucionan los problemas derivados de incompletitud de la información. Además, para agilizar el cálculo de la matriz de datos se seleccionan las 5 estaciones de tráfico más cercanas a cada estación de medición de la calidad del aire y el resto de estaciones se eliminan.

Homogeneización de los datos

Los sensores de tráfico registran mediciones cada 15 minutos, sin embargo a veces se producen retrasos y otros problemas que provocan que las horas de las mediciones en cada estación y cada día sean prácticamente aleatorias. Esto supone un problema para la combinación de los datos ya que es necesario tener un dato cada hora. Por otra, parte, aunque el periodo de integración de los datos es de 15 minutos, se producen muchas interrupciones en el funcionamiento de los sensores de tráfico. Esto provoca que haya periodos en los que un punto de medida de tráfico no

registra mediciones. Para solucionar estos dos problemas se homogeneizan los datos, empezando por la agrupación de datos a nivel de estación de tráfico. Dentro de una misma estación, para cada hora que se quiere utilizar en el modelo, se mira a ver si hay una medición registrada a la hora en punto. Es decir, se comprueba si hay una medición a las 00:00, a las 01:00, así hasta las 23:00. Si alguna de las horas no tiene registrada ninguna medición a la hora en punto hay dos opciones. Si la medición inmediatamente anterior está a una distancia temporal menor de una hora, se usa esta medida. Si la medición inmediatamente anterior está a más de una hora, se le pone un valor nulo a esa hora, con la finalidad de que sea eliminada posteriormente (ver Sección 3.1.2). Hay que destacar dos ventajas de esta solución. La primera es que al finalizar este proceso, cada estación tiene una fila de la matriz de datos para cada hora de cada día, por lo tanto todas las estaciones tienen el mismo número de mediciones, aunque alguna de las mediciones pueda ser nula. La segunda ventaja radica en que la información que no se tiene o se toma a pasado o se elimina, lo que permitiría implementar este sistema en tiempo real.

Cálculo de las variables de carga de tráfico

Para predecir la concentración de NO_2 es útil conocer no solo la carga de tráfico del momento del que quieres predecir el contaminante, si no también conocer como era la carga de tráfico en momentos anteriores. Sin embargo, obviamente no se dispone de la información del tráfico en el momento en el que se quiere predecir. Se presentan dos soluciones para este problema:

- Utilizar medidas reales de la carga de tráfico a pasado. Si se quiere predecir el nivel de concentración en tiempo t con un horizonte temporal de h horas el último dato de tráfico que se puede utilizar es el que se mide en tiempo $t - h$.
- Construir un modelo predictivo del tráfico en función del tráfico anterior y la metereología, de manera que se disponga de una estimación de la carga del tráfico para el momento en el que se quiere predecir la concentración de NO_2.

Se han implementado ambas soluciones, y ,por simplicidad, los resultados mostrados se hacen utilizando el tráfico a pasado. No obstante, en el anexo C se explica el modelo predictivo de tráfico y los resultados obtenidos utilizando dicho modelo.

Una vez que se toma la decisión de utilizar medidas reales del tráfico a pasado, se calculan nuevos atributos que se corresponden con la carga de tráfico en instantes de tiempo anteriores. La nomenclatura que se usa para estas nuevas variables es $carga\_nNsS$, que significa la variable de *carga* desplazada N intervalos de S horas hacia atrás en el tiempo. Además debe tener en cuenta el horizonte de predicción h que en nuestro caso son 24 horas. Es decir, en una fila cuya fecha indique el tiempo t (es decir sirve para predecir la concentración del contaminante en el tiempo t), la variable $carga\_nNsS$ contiene la carga que había en el momento $t - NS - h$ De esta manera, $carga\_n24s1$ y $carga\_n1s24$ son equivalentes y contienen la medición de la carga de tráfico de la misma estación dos días antes de la hora que se indica en la fila (la hora en la que se quiere predecir la concentración de NO_2).

Matriz final de datos de tráfico

En la matriz resultante del proceso al que se someten los datos de tráfico se tienen los siguientes campos: *idelem*, identificador único de la estación de medición del tráfico, $fecha$ *timestamp* de la medición, *carga*, la carga de tráfico y $carga\_nNsS$, valor de la carga en la misma estación desplazado N veces intervalos de S horas atrás en el tiempo teniendo en cuenta

Nombre	**Descripción**	**Nomenclatura**
idelem	identificador de la estación de medida	
fecha	'Timestamp' de la medición, contiene la fecha y hora.	t
carga	Medida del tráfico de 0 a 100 utilizada.	$c(t)$
carga_n1s1	Carga desplazada hacía atrás en el tiempo un día(h) y una hora.	$c(t-h-s)$
carga_n2s1	Carga desplazada hacia atrás en el tiempo un día(h) y dos horas.	$c(t-h-2s)$
carga_n3s1	Carga desplazada hacia atrás en el tiempo un día(h) 3 horas.	$c(t-h-3s)$
carga_n6s1	Carga desplazada hacia atrás en el tiempo un día(h) 6 horas.	$c(t-h-6s)$
carga_n1s24	Carga desplazada hacia atrás en el tiempo 2 días.	$c(t-h-24s)$

Tabla 3.1: Campos de los datos del tráfico procesados.

el horizonte h. Para entender mejor la estructura final de la matriz de datos de tráfico consúltese la Tabla 3.1.

La matriz resultante simplificada, siguiendo la nomenclatura introducida al principio, sería la siguiente:

$$\left[\begin{array}{cccccc|c}
c_1(t-n-h) & c_1(t-n-h-s) & c_1(t-n-h-2s) & c_1(t-n-h-3s) & c_1(t-n-h-6s) & c_1(t-n-h-24s) & t-n \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
c_1(t-h) & c_1(t-h-s) & c_1(t-h-2s) & c_1(t-h-3s) & c_1(t-h-6s) & c_1(t-h-24s) & t \\
c_2(t-n-h) & c_2(t-n-h-s) & c_2(t-n-h-2s) & c_2(t-n-h-3s) & c_2(t-n-h-6s) & c_2(t-n-h-24s) & t-n \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
c_2(t-h) & c_2(t-h-s) & c_2(t-h-2s) & c_2(t-h-3s) & c_2(t-h-6s) & c_2(t-h-24s) & t \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
c_K(t-n-h) & c_K(t-n-h-s) & c_K(t-n-h-2s) & c_K(t-n-h-3s) & c_K(t-n-h-6s) & c_K(t-n-h-24s) & t-n \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
c_K(t-h) & c_K(t-h-s) & c_K(t-h-2s) & c_K(t-h-3s) & c_K(t-h-6s) & c_K(t-h-24s) & t
\end{array}\right],$$

donde la última columna representa el *timestamp* y no se incluye en el modelo, sólamente se representa para una mejor comprensión de la matriz.

3.0.2. Datos de calidad del aire

Los datos de calidad del aire también son tratados antes de poder combinarse con el resto. El principal problema con el formato que se utiliza para guardar los datos es que en la matriz inicial todas las medidas de un mismo día están en una misma fila, por tanto para que el formato sea homogéneo con el resto de datos utilizados es necesario una transformación en la que cada medida tenga una fila distinta. Conviene recordar que para el módelo también se utilizarán los valores de concentración del contaminante en instantes de tiempo anteriores.

Transformación y limpieza de los datos

En primer lugar se hace una transformación de los datos, los datos horarios tienen el formato explicado en la Sección 2.4, se dispone de de una fila para todas las medidas horarias correspondientes a un único día. Para poder unir estos datos a los del tráfico y a los de metereología es necesario tener una fila para cada hora. Por lo tanto el primer paso es pasar de fila por cada día a 24 filas, una para cada hora.

Una vez que se tiene una medición horaria en cada fila, se eliminan aquellas filas que tengan en la columna de validación un valor distinto de 'V', es decir aquellas en las que la medición es incorrecta. Estas medidas eliminadas no se utilizarán para el modelo ya que no tiene sentido entrenarlo con datos erróneos. Además, para simplificar la construcción del modelo se mantienen solo las mediciones de NO_2 y se eliminan las correspondientes a otros contaminantes.

Nombre	Descripción
station_code	Código identificador de la estación
date	*Timestamp*
hour_val	Valor de concentración de NO_2 en el momento indicado por la columna *date*
hour_val_n0s1	Valor de concentración de NO_2 desplazado 24 horas hacia atrás en el tiempo.
hour_val_n1s1	Valor de concentración de NO_2 desplazado 25 horas hacia atrás en el tiempo.
hour_val_n2s1	Valor de concentración de NO_2 desplazado 26 horas hacia atrás en el tiempo.
hour_val_n3s1	Valor de concentración de NO_2 desplazado 27 horas hacia atrás en el tiempo.
hour_val_n4s1	Valor de concentración de NO_2 desplazado 28 horas hacia atrás en el tiempo.
hour_val_n1s24	Valor de concentración de NO_2 desplazado 48 horas hacia atrás en el tiempo.
hour_val_n2s24	Valor de concentración de NO_2 desplazado 72 horas hacia atrás en el tiempo

Tabla 3.2: Campos de los datos de calidad del aire procesados

Valor del contaminante a pasado

Al igual que con los datos de tráfico, es útil conocer el valor de la concentración de contaminate a pasado para predecirlo en un futuro. Con los datos de calidad del aire se generan nuevas columnas que son la columna de concentración de NO_2, llamada *hour_val*, desplazadas atrás en el tiempo. Al igual que con el tráfico si se quiere predecir la concentración del contaminante con un horizonte de un día es necesario que el último dato que se use para entrenar sea de al menos un día anterior. En otras palabras, la variable *hour_val* contiene la concentración de NO_2 en el momento de tiempo indicado por columna *date* (t), las columnas llamadas $hour\_val\_nNsS$ contienen la concentración de NO_2 en el momento $t-NS-h$ siendo h el horizonte de predicción. La Figura 3.1 muestra un ejemplo en el que el horizonte de predicción h es de 24 horas. En este caso h es un día, por tanto la columna $hour\_val\_n2s6$ contiene la concentración de NO_2 en el momento $t - 2 \times 6 - 24 = t - 36$, es decir, 36 horas antes del tiempo indicado en la columna *date*.

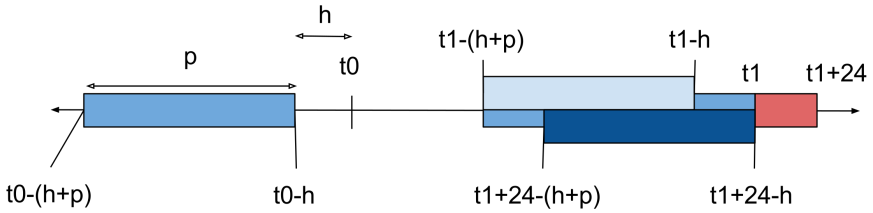


Figura 3.1: Línea temporal

Para hacer una predicción en $t0$ se necesita desde $t0 - (h + p)$ hasta $t0 - h$, este periodo se marca con un rectángulo azul. De igual manera, para hacer una predicción de todo un día empezando en $t1$, es decir, desde $t1$ hasta $t1 + 24$ hace falta la información desde $t1 - (h + p)$ hasta $t1$. El rectángulo de color azul claro representa el periodo necesario para predecir en $t1$, mientras que el rectángulo de más oscuro representa el periodo para predecir en $t1 + 24$

Los campos de la matriz de datos procesados están recogidos en la Tabla 3.2 y la matriz de datos de contaminación quedaría como sigue.

En la matriz de datos simplificada el resultado, usando las mismas variables de contaminación que se muestran en la Tabla 3.2 sería el siguiente,

$$\begin{bmatrix}
C_1(t_0) & C_1(t_0-24) & C_1(t_0-25) & C_1(t_0-26) & C_1(t_0-27) & C_1(t_0-28) & C_1(t_0-48) & C_1(t_0-72) \\
C_1(t_0+1) & C_1(t_0+1-24) & C_1(t_0+1-25) & C_1(t_0+1-26) & C_1(t_0+1-27) & C_1(t_0+1-28) & C_1(t_0+1-48) & C_1(t_0+1-72) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
C_1(t_0+23) & C_1(t_0+23-24) & C_1(t_0+23-25) & C_1(t_0+23-26) & C_1(t_0+23-27) & C_1(t_0+23-28) & C_1(t_0+23-48) & C_1(t_0+23-72) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
C_1(t_n) & C_1(t_n-24) & C_1(t_n-25) & C_1(t_n-26) & C_1(t_n-27) & C_1(t_n-28) & C_1(t_n-48) & C_1(t_n-72) \\
C_1(t_n+1) & C_1(t_n+1-24) & C_1(t_n+1-25) & C_1(t_n+1-26) & C_1(t_n+1-27) & C_1(t_n+1-28) & C_1(t_n+1-48) & C_1(t_n+1-72) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
C_1(t_n+23) & C_1(t_n+23-24) & C_1(t_n+23-25) & C_1(t_n+23-26) & C_1(t_n+23-27) & C_1(t_n+23-28) & C_1(t_n+23-48) & C_1(t_n+23-72) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
C_L(t_0) & C_L(t_0-24) & C_L(t_0-25) & C_L(t_0-26) & C_L(t_0-27) & C_L(t_0-28) & C_L(t_0-48) & C_L(t_0-72) \\
C_L(t_0+1) & C_L(t_0+1-24) & C_L(t_0+1-25) & C_L(t_0+1-26) & C_L(t_0+1-27) & C_L(t_0+1-28) & C_L(t_0+1-48) & C_L(t_0+1-72) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
C_L(t_0+23) & C_L(t_0+23-24) & C_L(t_0+23-25) & C_L(t_0+23-26) & C_L(t_0+23-27) & C_L(t_0+23-28) & C_L(t_0+23-48) & C_L(t_0+23-72) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
C_L(t_n) & C_L(t_n-24) & C_L(t_n-25) & C_L(t_n-26) & C_L(t_n-27) & C_L(t_n-28) & C_L(t_n-48) & C_L(t_n-72) \\
C_L(t_n+1) & C_L(t_n+1-24) & C_L(t_n+1-25) & C_L(t_n+1-26) & C_L(t_n+1-27) & C_L(t_n+1-28) & C_L(t_n+1-48) & C_L(t_n+1-72) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
C_L(t_n+23) & C_L(t_n+23-24) & C_L(t_n+23-25) & C_L(t_n+23-26) & C_L(t_n+23-27) & C_L(t_n+23-28) & C_L(t_n+23-48) & C_L(t_n+23-72)
\end{bmatrix},$$

siendo t_j siendo el tiempo $t-j$.

3.0.3. Datos de metereología

Los datos de metereología son tratados y procesados antes de unirlos con datos de tráfico y de calidad del aire siguiendo los pasos que se describen en los siguientes apartados.

Homegeneización de los datos

Los datos iniciales de predicciones metereológicas, cuyos campos se muestran en la Tabla 2.4, tienen una resolución de 3h. En otras palabras, sólo se tienen predicciones para las horas: 00:00, 03:00, 06:00, 09:00, 12:00, 15:00, 18:00 y 21:00 de cada día. Para completar las horas que faltan se hace una interpolación lineal entre los dos puntos más cercanos. Es decir, si definimos $v(t)$ como la predicción en el tiempo t entonces,

$$v(t+i) = \frac{(1-i)}{3}v(t) + \frac{i}{(3)}v(t+3) \; para \; i = 0,1,2,3, \;\; t = 0,3,6,9,...,3k, k \in \mathbb{N}$$

De manera que partiendo de la matriz

$$\begin{bmatrix}
v_1^1(0) & v_1^2(0) & \dots & v_d^1(0) & v_d^2(0) \\
v_1^1(3) & v_1^2(3) & \dots & v_d^1(3) & v_d^2(3) \\
\vdots & \vdots & \ddots & \vdots & \\
v_1^1(3k-3) & v_1^2(3k-3) & \dots & v_d^1(3k-3) & v_d^2(3k-3) \\
v_1^1(3k) & v_1^2(3k) & \dots & v_d^1(3k) & v_d^2(3k)
\end{bmatrix}$$

se obtiene a la matriz:

$$\begin{bmatrix}
v_1^1(0) & v_1^2(0) & \dots & v_d^1(0) & v_d^2(0) \\
v_1^1(1) & v_1^2(1) & \dots & v_d^1(1) & v_d^2(1) \\
v_1^1(2) & v_1^2(2) & \dots & v_d^1(2) & v_d^2(2) \\
v_1^1(3) & v_1^2(3) & \dots & v_d^1(3) & v_d^2(3) \\
\vdots & \vdots & \ddots & \vdots & \\
v_1^1(3k-1) & v_1^2(3k-1) & \dots & v_d^1(3k-1) & v_d^2(3k-1) \\
v_1^1(3k) & v_1^2(3k) & \dots & v_d^1(3k) & v_d^2(3k)
\end{bmatrix}$$

Además en vez de usar las componentes horizontal y vertical del viento se cálcula el módulo para quedárse sólo con la información correspondiente a la velocidad del viento, sin tener en cuenta la dirección ni el sentido.

Cálculo de las variables metereológicas a pasado

Es de gran utilidad conocer la metereología no sólamente en el momento en el que se quiere predecir si no también en momentos anteriores. Para tener esta información en la matriz de datos se cálculan nuevas columnas que contienen las variables metereológicas con valores pasados. Por cada una de las variables metereológicas: temperatura, humedad relativa, velocidad del viento y presión atmosférica se calculan variables que representan sus valores a pasado. En concreto, si t es el timestamp de una fila, se calculan: t-3, t-6 y t-9 de forma que podemos definir

$$\hat{M}_j(t,3) = \{v^j_{temp}(t), v^j_{temp}(t-s), v^j_{temp}(t-2s), v^j_{temp}(t-3s), \ldots, v^j_{viento}(t), v^j_{viento}(t-s), v^j_{viento}(t-2s), v^j_{viento}(t-3s)\}$$

En este caso se ha elegido $s = 3$ por tanto:

$$\hat{M}_j(t,3) = \}v^j_{temp}(t), v^j_{temp}(t-3), v^j_{temp}(t-6), v^j_{temp}(t-9), \ldots, v^j_{vienot}(t), v^j_{viento}(t-3), v^j_{viento}(t-6), v^j_{viento}(t-9)\{$$

Y la matriz resultante por lo tanto es la siguiente,

$$\begin{bmatrix} \hat{M}_1(t-n) & \hat{M}_2(t-n) \\ \vdots & \vdots \\ \hat{M}_1(t) & \hat{M}_2(t) \end{bmatrix},$$

donde t representa el último tiempo en el que queremos predecir y $t-k$ es el momento correspondiente a k horas antes de t.

3.1. Alineación espacial y temporal

Para poder combinar los distintos datos entre sí, además de procesarlos previamente hay que alinearlos tanto temporal como espacialmente. Se necesita que las mediciones en la calidad del aire se combinen con las mediciones de tráfico y de metereología que se han tomado a la misma hora. Además no es informativo utilizar la información del tráfico de un punto que está muy lejos de la estación de calidad del aire en la que se quiere predecir la concentración de NO_2, y,por tanto hay que tener en cuenta la distancia entre estaciones de tráfico y estaciones de calidad del aire.

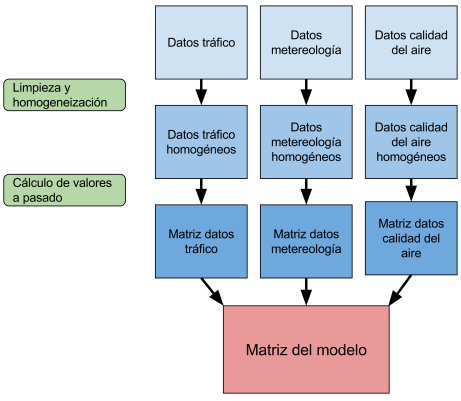


Figura 3.2: Esquema de desarrollo

3.1.1. Sincronización temporal

Para la sincronización temporal la parte principal se hace en la parte de procesamiento de los datos, ya que se homegeneizan los datos, poniendo una medida cada hora en todos los conjuntos de datos: tráfico, metereología y calidad del aire. Una vez que se tiene una medida cada hora solamente hay que combinar las tablas utilizando las columnas de *timestamp* como claves. De esta manera se consigue una nueva matriz de datos combinando la matriz de metereología y la de calidad del aire, y a la que llamaremos matriz secundaria. El resultado es una matriz que tiene las mismas columnas que la matriz de calidad del aire (Tabla 3.2) y además contiene las columnas con la información de las predicciones metereológicas a la hora en la que se tomó la medida de calidad del aire. Para combinar esta nueva matriz con los datos del tráfico es necesario llevar a cabo otro proceso que se explica en el siguiente punto.

3.1.2. Alineación espacial

Para la sincronización espacial se ha desarrollado un módulo de distancias cuya finalidad es calcular la distancia entre un punto de medida de tráfico y una estación de calidad del aire, así como calcular las estaciones de tráfico más cercanas a una estación de contaminación dada. Este módulo se utiliza para la combinación de los datos del tráfico con el resto. El proceso es el siguiente: para cada estación de calidad del aire se seleccionan las K estaciones de tráfico más cercanas, de forma que a la matriz secundaria se le añaden K nuevas columnas, que son el tráfico en cada una de las K estaciones en la fecha y hora indicadas por la fila de la matriz secundaria.

Se tienen en un principio dos matrices,

$$\begin{bmatrix} c_1(t-n) & c_1(t-n-s) & c_1(t-n-2s) & c_1(t-n-3s) & c_1(t-n-6s) & c_1(t-n-24s) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_1(t) & c_1(t-s) & c_1(t-2s) & c_1(t-3s) & c_1(t-6s) & c_1(t-24s) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_K(t-n) & c_K(t-n-s) & c_K(t-n-2s) & c_K(t-n-3s) & c_K(t-n-6s) & c_K(t-n-24s) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_K(t) & c_K(t-s) & c_K(t-2s) & c_K(t-3s) & c_K(t-6s) & c_K(t-24s) \end{bmatrix}$$

y

$$\left[\begin{array}{ccc|c} C(t_0-(n+h)s) & \hat{M}_1(t_0-ns) & \hat{M}_2(t_0-ns) & C(t_0-ns) \\ C(t_0-(n-1+h)s) & \hat{M}_1(t_0-(n-1)s) & \hat{M}_2(t_0-(n-1)s) & C(t_0-(n-1)s) \\ \vdots & \vdots & \vdots & \vdots \\ C(t_0-(1+h)s) & \hat{M}_1(t_0-s) & \hat{M}_2(t_0-s) & C(t_0-s) \\ C(t_0-hs) & \hat{M}_1(t_0) & \hat{M}_2(t_0) & C(t_0) \end{array}\right]$$

Estas dos matrices se combinan de manera que la matriz resultante es la que se muestra a continuación. Hay que notar que en la matriz resultante que servirá de entrada para los métodos de aprendizaje se eliminan todas aquellas filas en las que alguna variable tenga un valor nulo.

$$\left[\begin{array}{cccccc|c} C(t_0-(n+h)s) & \hat{M}_1(t_0-ns) & \hat{M}_2(t_0-ns) & c_1(t_0-(n+h)s) & \dots & c_K(t_0-(n+h)s) & C(t_0-ns) \\ C(t_0-(n-1+h)s) & \hat{M}_1(t_0-(n-1)s) & \hat{M}_2(t_0-(n-1)s) & c_1(t_0-(n-1+h)s) & \dots & c_K(t_0-(n-1+h)s) & C(t_0-(n-1)s) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ C(t_0-(1+h)s) & \hat{M}_1(t_0-s) & \hat{M}_2(t_0-s) & c_1(t_0-s) & \dots & c_K(t_0-s) & C(t_0-s) \\ C(t_0-hs) & \hat{M}_1(t_0) & \hat{M}_2(t_0) & c_1(t_0) & \dots & c_K(t_0) & C(t_0) \end{array}\right]$$

4

Análisis descriptivo y resultados

4.1. Introducción

En este capítulo se exponen y analizan los resultados obtenidos tanto en la predicción de la concentración de NO_2 como en el análisis descriptivo de los datos utilizados. La primera sección se centra en describir la implementación que se ha llevado a cabo. Mientras que las dos siguientes secciones mostrarán el análisis descriptivo de los datos y los resultados obtenidos, respectivamente.

4.2. Implementación

Debido a la complejidad del proceso de transformación y combinación de los datos (Sección **??**, y de cara a poder realizar operaciones sobre las mismas como calcular distancias rápidamente o aplicar algoritmos de aprendizaje automático, se han utilizado diversas librerías de Python, entre las que cabe destacar las siguientes:

- **Scipy**: No es un paquete en sí, si no que se trata de un ecosistema que contiene los paquetes que se han usado principalmente, tales como numpy, scikit-learn, pandas y matplotlib. Debido a la importancia de estos paquetes, se explicará cada uno por separado [25].
- **Numpy**: Numpy es el paquete principal para computación científica en Python, ofreciendo, entre otras cosas, un potente array N-dimensional, implementaciones de funciones sofisticadas o herramientas para integrarse con C o Fortran. En este trabajo numpy se ha usado durante todo el desarrollo, especialmente en la sección de aprendizaje automático [26].
- **Pandas**: Se trata de una librería de código abierto que provee eficientes estructuras de datos y herramientas de análisis para Python. La principal ventaja sobre numpy es que ofrece herramientas de alto nivel de manipulación de datos construidas sobre numpy, lo que permite un mayor nivel de abstracción. Toda la parte de tratamiento y transformación de los datos se ha realizado utilizando Pandas [27].

- **scikit-learn**: Es una libreria de código abierto que implementa un amplio rango de algoritmos de aprendizaje automático, validación cruzada, preprocesamiento y visualización utilizando una interfaz unificada. En este trabajo se ha usado para todos los algoritmos de aprendizaje automático considerados [28].
- **Matplotlib**: Es una librería del entorno de Scipy, también introducida dentro de scikit-learn, numpy o pandas que se utiliza para hacer gráficas [29].

4.3. Análisis descriptivo

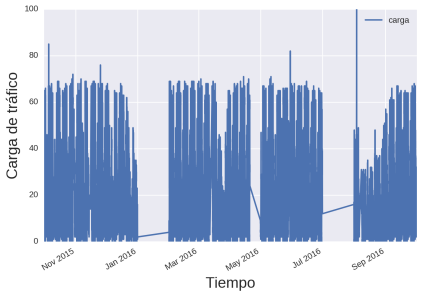
4.3.1. Análisis de los datos del tráfico

Para analizar los datos del tráfico se ha usado el promedio de los datos registrados en distintos puntos de medida durante un año, desde el 01/10/2015 al 01/10/2016. Es decir, no se utiliza un único punto de medida de la carga de tráfico si no que, en esta sección, se analizan todos en conjunto. Con el objetivo de analizar los datos se ha generado la gráfica de la serie temporal de la carga media del trabajo (Figura C.2a) y se han creado diagramas de caja agrupados por el tiempo y en tres escalas diferentes: diaria (Figura C.2b), semanal(Figura C.2c) y anual(Figura C.2d). En la escala semanal cada día de la semana se representa con un número entero, siendo el lunes el 0 y el domingo el 6. En esta gráfica de datos diarios se pueden ver importantes diferencias entre el tráfico durante el día y durante la noche, incluso se pueden notar las horas puntas que se producen a primera hora de la mañana y por la tarde.

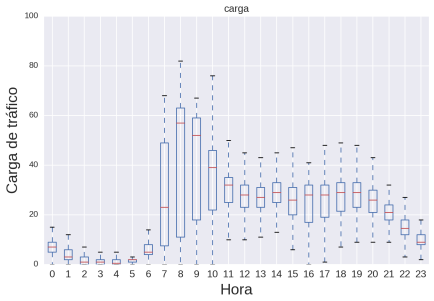
Los diagramas de caja además confirman hechos esperados, a nivel semanal se comprueba por ejemplo que la carga de tráfico es menor los fines de semana, mientras que a nivel mensual se puede observar que el tráfico en Madrid disminuye en agosto, ya que la mayoría de la gente sale de vacaciones.

Autocorrelación del tráfico

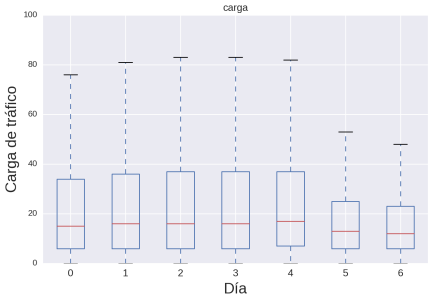
Con el objetivo de determinar qué relación hay entre el tráfico en el momento de la predicción y el tráfico en tiempos anteriores, que es lo que utiliza el modelo, se analiza la autocorrelación de la carga del tráfico. Para ello, se representa la carga del tráfico como una serie temporal y se mide su autocorrelación, es decir la correlación del valor de la serie temporal con la misma serie desplazada en el tiempo a pasado una determinada cantidad h. En la Figura 4.2 se puede observar que la autocorrelación de la carga de tráfico con un desplazamiento de 24 horas es muy elevada. Para medir la correlación se ha utilizado el promedio de los datos en todas las estaciones de tráfico durante el periodo de tiempo transcurrido entre el 01/02/2016 y 01/04/2016.



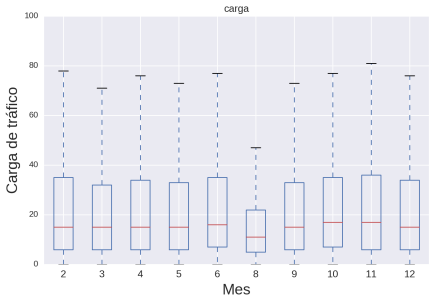
(a) Gráfica de la serie temporal del tráfico. Se puede observar que no se toman mediciones en enero ni julio.



(b) Diagrama de cajas del tráfico diario. Se observa como aumenta la carga en las horas punta mientras que disminuye considerablemente por la noche

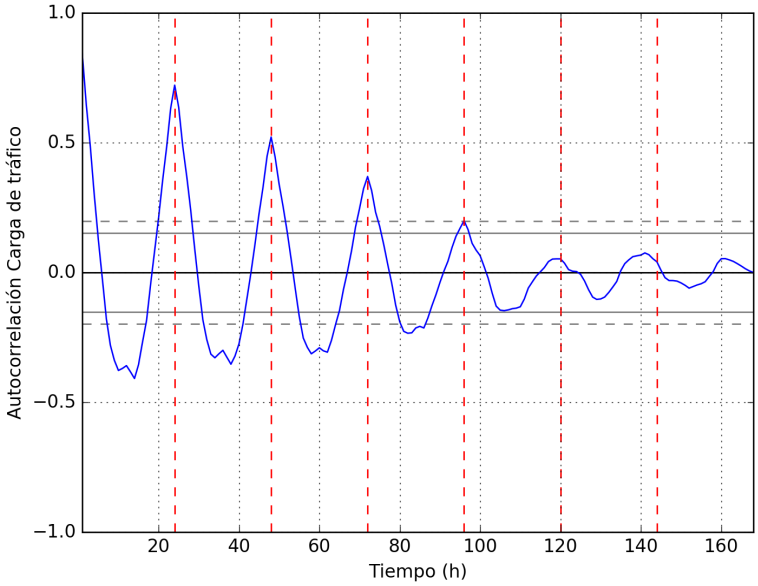


(c) Diagrama de cajas del tráfico semanal, se puede apreciar la disminución de la carga los fines de semana.

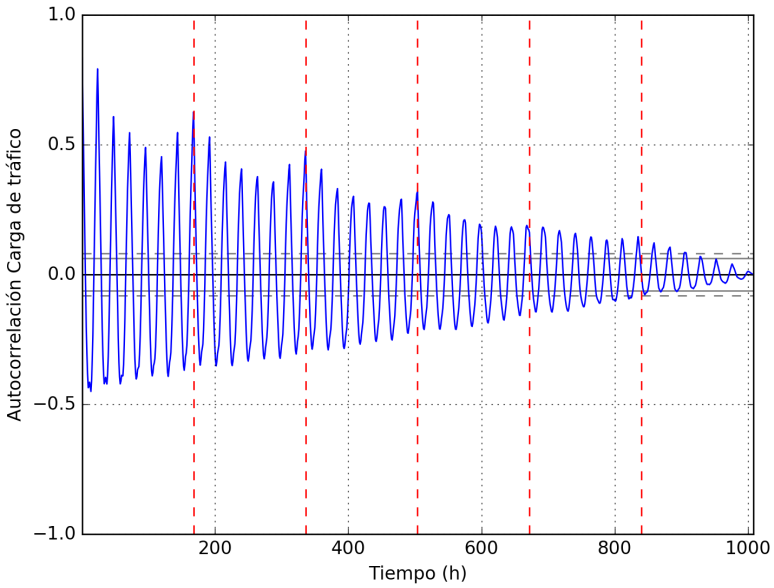


(d) Diagrama de cajas del tráfico mensual. No hay mediciones de enero ni julio. Se puede observar que la menor carga se tiene en el mes de Agosto.

Figura 4.1: Análisis de la carga de tráfico en promedio en Madrid en el periodo de tiempo de un año transcurrido entre 01/10/2015 y 01/10/2016.



(a) Gráfica de autocorrelación del tráfico hasta una semana. Se muestran líneas rojas verticales en los múltiplos de 24 horas. Se observa que se obtienen máximos locales cada 24 horas, disminuyendo estos máximos en magnitud según transcurre el tiempo. Para mantener la distancia temporal natural de siete días entre cada semana no se han diferenciado los días entre semana del sábado y el domingo, lo que puede disminuir la autocorrelación real.



(b) Gráfica de autocorrelación del tráfico hasta 6 semanas. Se muestran líneas rojas verticales cada 7 días. Se puede observar que cada semana se vuelve a producir una autocorrelación mayor que en los tiempos más cercanos que no corresponden a semanas enteras. Para mantener la distancia temporal natural de siete días entre cada semana no se han diferenciado los días entre semana del sábado y el domingo, lo que puede disminuir la autocorrelación real.

Figura 4.2: Gráficas de autocorrelación de la carga de tráfico promedio en Madrid utilizando dos meses de datos, desde el 01/02/2016 al 01/04/2016.

4.3.2. Análisis de los datos de calidad del aire

Para el análisis de los datos de calidad del aire se han considerado un subconjunto de estaciones que se encuentran en localizaciones céntricas. Las estaciones elegidas son: Plaza de España, Escuelas Aguirre, Méndez Álvaro y Plaza del Carmen. Sus identificadores y direcciones se muestran en la Tabla 4.1

Idelem	**Nombre**	**Dirección**
28079004	Plaza de España	C/ Princesa esq. Plaza de España
28079008	Escuelas Aguirre	Entre c/ Alcalá y c/ O'Donnell
28079047	Méndez Álvaro	C/Juan de Mariana - Pza. Amanecer Mendez Alvaro
28079035	Plaza del Carmen	Plaza del Carmen esq. Tres Cruces

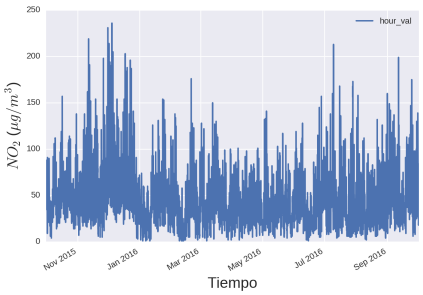
Tabla 4.1: Estaciones de calidad del aire elegidas para el análisis y resultados.

Al igual que con los datos de tráfico también se ha usado un año de datos de calidad del aire, correspondiente al periodo entre el 01/10/2015 y el 01/10/2016. Las figuras 4.3b, 4.3c y 4.3d muestran los diagramas de caja de la concentración de NO_2 para distintas divisiones temporales (diaria, semanal y mensual). La figura 4.3a representa la concentración de NO_2 con respecto al tiempo. En este caso los resultados se han calculado individualmente para cada estación de calidad del aire seleccionada. Se puede hacer un análisis de los diagramas de cajas similar el del tráfico ya que, la concentración de NO_2 baja los fines de semana(4.3c) y sube en horas punta de la mañana (4.3b), incluso el valor más bajo se obtiene en agosto (4.3d). Estas similitudes con la carga de tráfico invitan a pensar que la concentración de NO_2 es, efectivamente, una consecuencia directa del tráfico. Sin embargo, es destacable remarcar que sin razón aparente la concentración de NO_2 es mayor por ejemplo en noviembre que en febrero, lo que no puede ser explicado con diferencias en el tráfico. Quizás esto se deba a dependencias con la metereología ya que los meses de septiembre y octubre tuvieron menos precipitaciones que diciembre y enero.

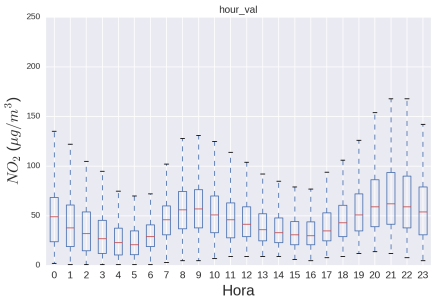
Autocorrelación de la concentración de NO_2

Con el objetivo de determinar posibles autocorrelaciones en la señal de NO_2, se han realizado gráficas de autorrelación similares a las generadas para la carga de tráfico.

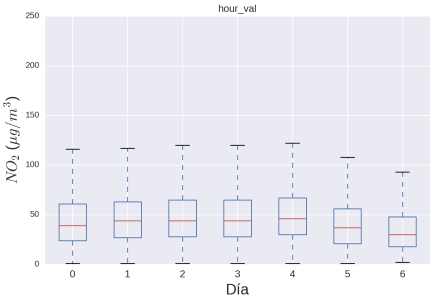
En la Figura 4.4 se puede observar que las líneas verticales que representan los días enteros en la figura 4.4a y las semanas enteras en la figura 4.4b coinciden con máximos locales en la autocorrelación. Sin embargo, el valor de la autocorrelación en un día que es 0.24 indica que un modelo lineal que sólo utilizase valores de la concentración de NO_2 a pasado no obtendría buenos resultados, esto es algo que puede comprobarse con los resultados que se presentarán en la Sección 4.4.



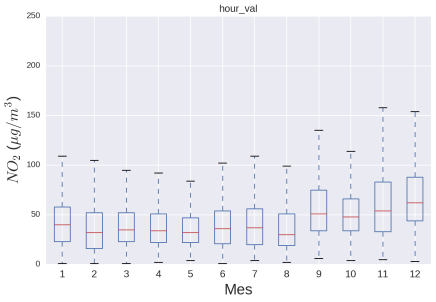
(a) Gráfica de la serie temporal de la concentración de NO_2. Se puede observar que no se toman mediciones en enero ni julio.



(b) Diagrama de cajas de la concentración de NO_2 diaria. Se observa como aumenta la carga en las horas punta mientras que disminuye considerablemente por la noche

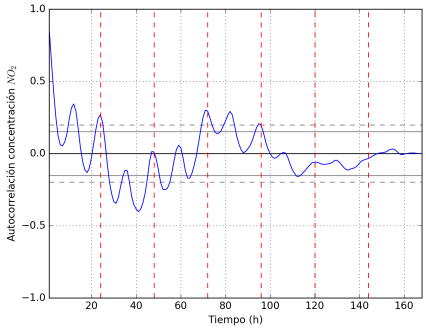


(c) Diagrama de cajas de la concentración de NO_2 semanal, se puede apreciar la disminución de la carga los fines de semana.

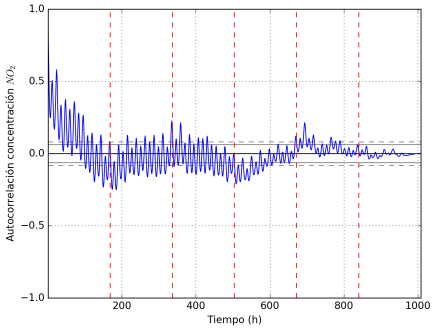


(d) Diagrama de cajas de la concentración de NO_2 mensual. No hay mediciones de enero ni julio. Se puede observar que en agosto es menor la carga.

Figura 4.3: Análisis de la concentración de NO_2 medido entre el 01/10/2015 y el 01/10/2016 en la estación de Plaza de España.

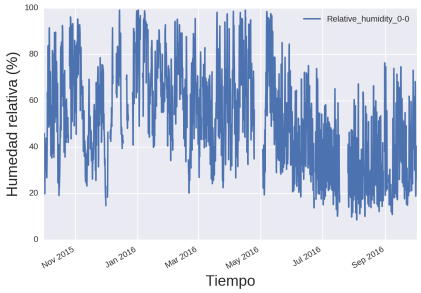


(a) Gráfica de autocorrelación de la concentración de NO_2 hasta una semana. Se muestran líneas rojas verticales en los múltiplos de 24 horas. Se observa que se obtienen máximos cada 24 horas, alternándose estos máximos en magnitud según transcurre el tiempo.

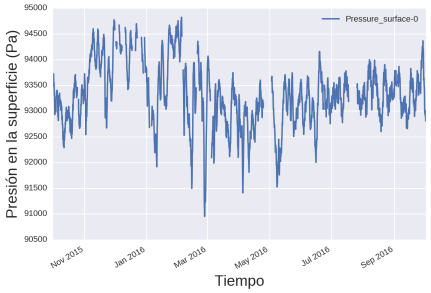


(b) Gráfica de autocorrelación de la concentración de NO_2 hasta 6 semanas. Se muestran líneas rojas verticales cada 7 días.

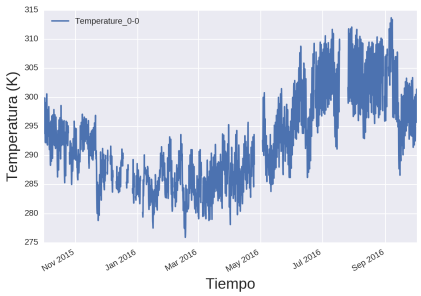
Figura 4.4: Gráficas de autocorrelación de la concentración de NO_2 medido como promedio en todas las estaciones consideradas y teniendo en cuenta el periodo de tiempo transcurrido entre el 01/02/2016 y el 01/04/2016.



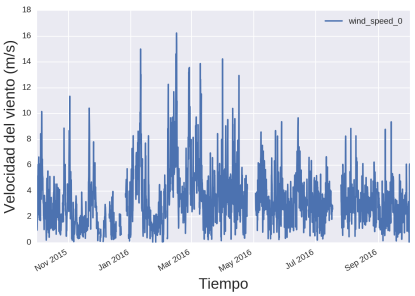
(a) Gráfica de la serie temporal de la humedad relativa medida como un porcentaje.



(b) Gráfica de la serie temporal de la presión en la superficie medida en Pascales.



(c) Gráfica de la serie temporal de la temperatura medida en Kelvin



(d) Gráfica de la serie temporal de la velocidad del viento medida en m/s.

Figura 4.5: Gráficas temporales que representan las variables metereológicas utilizadas para el modelo correspondientes al punto situado al oeste de Madrid.

4.3.3. Análisis de los datos del GFS

El análisis de los datos metereológicos se ha realizado a partir de las predicciones realizadas durante un año, desde el 01/10/2015 hasta el 01/10/2016. En la figura 4.5 se muestran las series temporales de las variables metereológicas utilizadas: temperatura, velocidad del viento, presión y humedad relativa.

4.3.4. Correlación entre variables

Los modelos lineales como la regresión lineal o el método Lasso explicados en la Sección 2.7 utilizan la correlación de Pearson entre las variables regresoras y la variable explicada para calcular los coeficientes que minimizan el error de regresión. La correlación entre variables representa la dependencia lineal entre ambas. Podemos estudiar la correlación entre las variables utilizadas para estudiar la dependencia entre ellas. Además, la correlación entre las variables regresoras y la variable explicada nos da una idea de cómo de buena es cada variable regresora para predecir. Para conocer qué variables aportan más información sobre la concentración de NO_2 y para saber qué variables están correlacionadas entre sí, se calcula la matriz de correlación y se representa como un mapa de calor. Las Figuras 4.6 y 4.7 muestran la correlacion entre las variables de carga del tráfico y la concentración de NO_2. Los mapas de calor se segregan por estaciones de calidad del aire, de esta manera se pueden relacionar con los resultados expuestos en la Sección 4.4.

Las variables de carga reciben una nueva nomenclatura ya que hay distintas estaciones. La

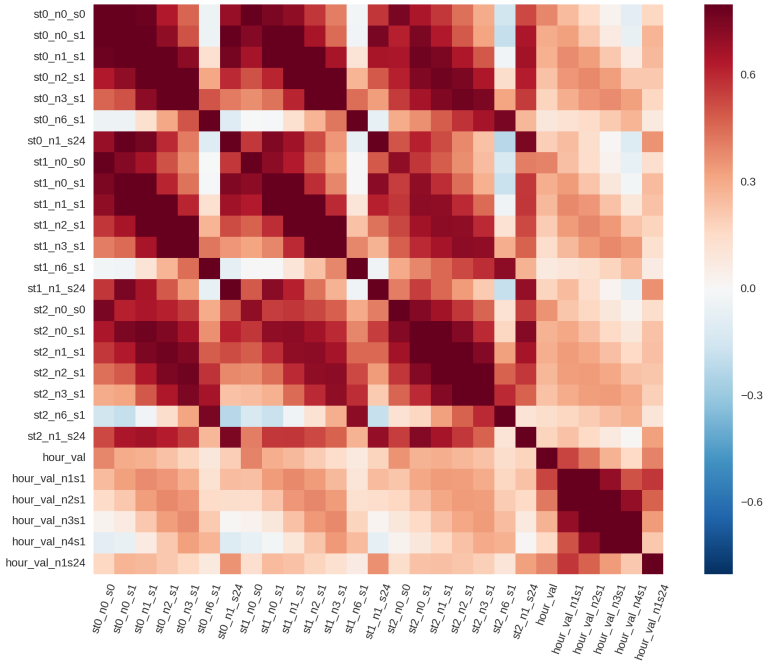


Figura 4.6: Mapa de calor de la correlación entre la concentración de NO_2 y la carga del tráfico en la estación Escuelas Aguirre.

notación $carga\_nNsS$ definida al inicio de la sección 3 define la variable que a la concentración de NO_2 en t (variable $hour\_val$) le asocia la carga de tráfico en $t - ns - h$. Para una mejor representación en los mapas de calor, en lugar de las 5 estaciones utilizadas en modelo, se han cogido las 3 estaciones de tráfico más cercanas a la estación de contaminación y se definen las variables stj_nNsS para $j = 0, 1, 2$ como la variable $carga\_nNsS$ medida en la estación j. Por ejemplo, $st0\_n2s1$ contiene la carga de tráfico en la estación 0 en el momento $t - 2 - 24$. De igual modo, es necesario definir que las variables denominadas $hour\_val\_nNsS$ son aquellas que a la concentración de NO_2 en t (variable $hour\_val$) le asocia la misma concentración de NO_2 en el instante $t - ns - h$. Por ejemplo, $hour\_val\_n2s1$ contiene la concentración de NO_2 en el momento $t - 2 - 24$.

En las Figuras 4.6 y 4.7 se pueden hacer varias apreciaciones. La primera es notar que las estaciones de tráfico están muy correladas entre sí, lo que, entre otras posibilidades, puede significar que son muy cercanas. La segunda observación es que la correlación entre la concentración de NO_2 y la carga de tráfico es en general mayor en la estación Escuelas Aguirre que en la Plaza del Carmen. Este resultado apoya los resultados obtenidos en la Sección 4.4, donde el rendimiento de los modelos obtenidos es mejor para la estación Escuelas Aguirre.

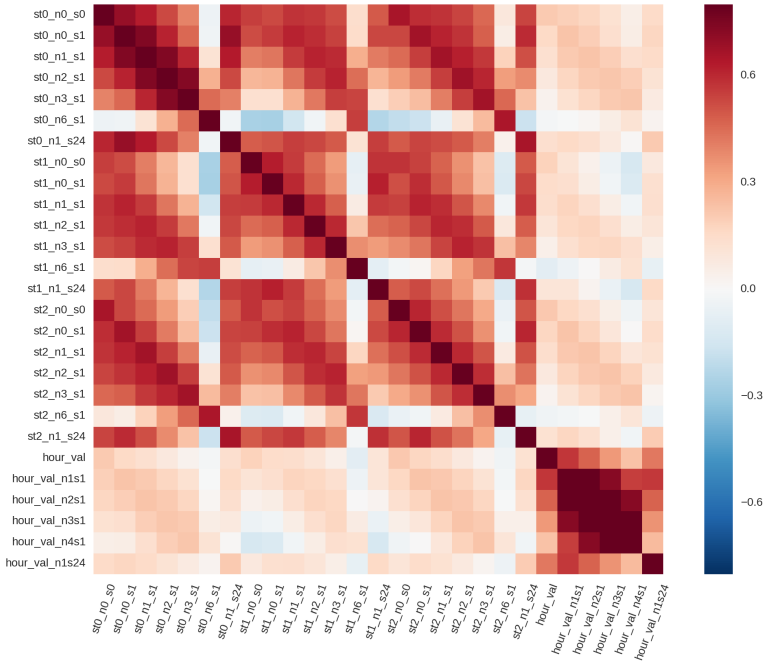


Figura 4.7: Mapa de calor de la correlación entre la concentración de NO_2 y la carga del tráfico en la estación Plaza del Carmen.

4.4. Modelo predictivo de contaminación

4.4.1. Resultados Regresión para la predicción de la concentración de NO_2

En este trabajo se pretende desarrollar un modelo que prediga la concentración horaria de NO_2 con un horizonte de predicción de 24 horas y utilizando datos históricos de concentración de NO_2, datos de tráfico y datos de predicciones metereológicas. Para analizar la importancia de los datos externos a la concentración de NO_2 se presentan los resultados de tres modelos distintos:

- **Modelo a pasado(Pasado)**: Es el modelo que sólo utiliza la información de la concentración del contaminante en momentos de al menos un día antes al momento en que se quiere predecir. Para generarlo no se utilizan datos del tráfico ni metereológicos.
- **Modelo con tráfico(Tráfico)**: Es el modelo que utiliza tanto el valor del contaminante a pasado como datos de la carga de tráfico a pasado en los 5 puntos de medida más cercanos a la estación de medida de concentración de NO_2. No utiliza datos meterológicos.
- **Modelo completo(Completo)**: Es el modelo completo que incluye datos de la concentración de contaminante a pasado, la carga del tráfico a pasado y las predicciones metereológicas.

Los resultados se disgregan por estación, es decir, se entrena un modelo para cada estación y los resultados se obtienen evaluando sobre datos de la misma estación. En los modelos de cada estación se utilizan datos de un año y 3 meses (del 01/10/2015 al 31/12/2016) utilizando un año para entrenamiento (del 01/10/2015 al 30/9/2016) y tres meses para test (del 01/10/2016 al 31/12/2016). Además, para cada modelo se han usado los algoritmos de regresión explicados en la sección 2.7: Random Forest de regresión (RandomForestRegressor), Lasso (Lasso), ElasticNet (ElasticNet) y Regresión Lineal (LinearRegression). El ajuste de los parámetros de estos algoritmos se realiza mediante el algoritmo GridSearch proporcionado por scikit-learn que realiza una búsqueda en una rejilla predefinida y realizando validación cruzada sobre el conjunto de entrenamiento. Para comparar los distintos modelos y algoritmos de regresión entre sí se calculan dos medidas. La primera es el coeficiente de determinación R^2, una medida de la bondad del ajuste en el modelo de regresión que se calcula como:

$$R^2 = \frac{SCR}{SCE} ,$$

donde $SCR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ mide la variabilidad explicada por el modelo y $SCE = \sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ mide la variabilidad no explicada por el modelo. Por lo tanto, se puede entender R^2 como la proporción de variabilidad que explica el modelo. La segunda medida utilizada es el $MAPE$ (Mean Absolute Percentage Error) que se calcula como:

$$MAPE = \sum_{i=1}^{n} \frac{|\hat{Y}_i - Y_i|}{|Y_i|} ,$$

y representa cuánto se equivoca el modelo de media medido como porcentaje del valor real. En las Tablas 4.2, 4.3, 4.4 y 4.5 se muestran los resultados obtenidos para las estaciones de la Tabla 4.1.

A partir de los resultados mostrados las Tablas 4.2, 4.3 , 4.4 y 4.5 se pueden extraer conclusiones sobre el modelo predictivo que se ha desarrollado. En primer lugar, la carga de tráfico tiene un efecto en la concentración de NO_2, ya que todos los algoritmos utilizados mejoran su

Estación: Plaza de España		**Pasado**	**Tráfico**	**Completo**
RandomForestRegressor	R^2	0.3189	0.4683	0.4124
	$MAPE$	24.42 %	**17.37 %**	23.69 %
Lasso ($\alpha = 0.495$)	R^2	0.3412	**0.5714**	0.5585
	$MAPE$	24.38 %	19.04 %	20.10 %
ElasticNet ($\alpha = 0.55$)	R^2	0.3184	0.5695	0.5602
	$MAPE$	23.34 %	19.00 %	20.07 %
LinearRegression	R^2	0.3235	0.5525	0.5438
	$MAPE$	24.27 %	19.48 %	20.31 %

Tabla 4.2: Resultados para la predicción de la concentración de NO_2 en la estación Plaza de España. Se muestra tanto el coeficiente R^2 como el $MAPE$ para cada uno de los cuatro modelos de regresión utilizados.

Estación: Escuelas Aguirre		**Pasado**	**Tráfico**	**Completo**
RandomForestRegressor	R^2	0.2802	0.4752	**0.5682**
	$MAPE$	23.99 %	23.14 %	23.72 %
Lasso ($\alpha = 1.115$)	R^2	0.3836	0.5083	0.5481
	$MAPE$	23.23 %	21.45 %	21.13 %
ElasticNet ($\alpha = 1.485$)	R^2	0.3422	0.5041	0.5491
	$MAPE$	21.62 %	21.33 %	21.11 %
LinearRegression	R^2	0.3644	0.5184	0.5568
	$MAPE$	23.43 %	21.44 %	**20.74 %**

Tabla 4.3: Resultados para la predicción de la concentración de NO_2 en la estación Escuelas Aguirre. Se muestra tanto el coeficiente R^2 como el $MAPE$ para cada uno de los cuatro modelos de regresión utilizados.

rendimiento al incluir este tipo de variables. Se puede establecer una relación inmediata entre este resultado y la correlación de la concentración de NO_2 con la carga de tráfico. Se puede observar que la correlación entre variables de carga de tráfico y concentración de NO_2 es mayor en Escuelas Aguirre (Figura 4.6) que en Plaza del Carmen (Figura 4.7). De igual manera, los rendimientos obtenidos en Escuelas Aguirre (Tabla 4.3) son superiores a los rendimientos en Plaza del Carmen (Tabla 4.5). Deducimos, por lo tanto, que las variables de carga de tráfico guardan una relación con la concentración de NO_2 y esta relación se expresa como modelos de

Estación: Méndez Álvaro		**Pasado**	**Tráfico**	**Completo**
RandomForestRegressor	R^2	0.3443	0.3745	0.5129
	$MAPE$	27.11 %	21.39 %	31.71 %
Lasso ($\alpha = 0.77$)	R^2	0.3904	0.4188	0.5214
	$MAPE$	23.77 %	22.05 %	30.22 %
ElasticNet ($\alpha = 0.84$)	R^2	0.3341	0.5041	0.5177
	$MAPE$	25.56 %	21.33 %	30.51 %
LinearRegression	R^2	0.3474	0.3535	**0.5229**
	$MAPE$	24.15 %	23.43 %	**19.51 %**

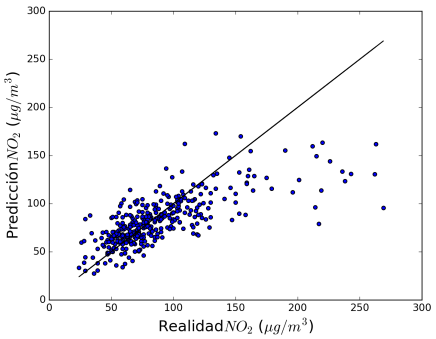
Tabla 4.4: Resultados para la predicción de la concentración de NO_2 en la estación Méndez Álvaro. Se muestra tanto el coeficiente R^2 como el $MAPE$ para cada uno de los cuatro modelos de regresión utilizados.

Estación: Plaza del Carmen		**Pasado**	**Tráfico**	**Completo**
RandomForestRegressor	R^2	0.2800	0.4794	0.2812
	$MAPE$	21.46 %	33.23 %	23.69 %
Lasso ($\alpha = 0.655$)	R^2	0.2994	0.4739	**0.5272**
	$MAPE$	20.23 %	34.63 %	**19.51 %**
ElasticNet ($\alpha = 0.32$)	R^2	0.3263	0.4777	0.5247
	$MAPE$	19.61 %	34.41 %	19.58 %
LinearRegression	R^2	0.3289	0.4545	0.4762
	$MAPE$	19.88 %	35.29 %	31.35 %

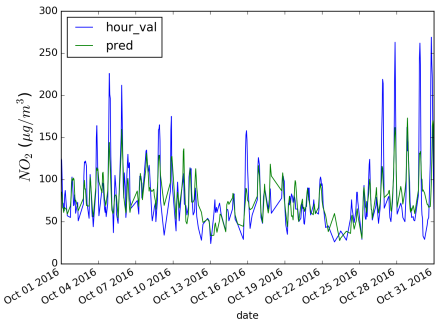
Tabla 4.5: Resultados para la predicción de la concentración de NO_2 en la estación Plaza del Carmen. Se muestra tanto el coeficiente R^2 como el $MAPE$ para cada uno de los cuatro modelos de regresión utilizados.

regresión de la concentración de NO_2 más precisos. Por otra parte, las predicciones metereológicas, aunque aportan cierta información útil, no parecen ser tan determinantes para la predicción de NO_2: la inclusión de variables metereológicas a veces conduce a un mejor rendimiento (Tabla 4.4) mientras que otras veces lleva a un rendimiento más pobre (Tabla 4.2). A pesar de que largos periodos sin precipitaciones suelen desencadenar etapas con alta concentración de NO_2, el modelo no refleja estas situaciones. Una posible solución a este problema, propuesta para trabajo futuro, es incluir nuevas variables al modelo que utilicen esta información, por ejemplo, tiempo transcurrido desde la última vez que llovió.

Para hacerse una idea visual del resultado de la predicción se generan gráficas en las que se representa la predicción contra la realidad. Si el modelo no tuviera error todos los puntos se colocarían sobre la recta bisectriz. Aunque los resultados obtenidos por los distintos métodos son parecidos, ElasticNet con las variables de meteorología y tráfico es, en media, el que obtiene los mejores resultados. Por ello se ha seleccionado las predicciones este método para generar las gráficas de la Figura 4.8. Se puede apreciar en esta figura que el modelo se centra en predecir correctamente los valores de concentración de NO_2 entre 50 y 150 ya que la mayoría de los valores de concentración se encuentran en esta franja, como se observa en la figura 4.3. El problema es que el modelo subestima de manera sistemática valores altos de concentración, lo que no es deseable si el objetivo es predecir las alertas de alta concentración de NO_2. Para intentar explicar este hecho se proponen dos hipótesis no excluyentes. La primera hipótesis se basa en el hecho de que la mayoría de datos de concentración de NO_2 corresponden a valores menores de 180 $\mu g/m^3$, y los modelos lineales, que se basan en reducir el error cuadrático medio, tienen tendencia a ajustarse mejor en las zonas con mayor concentración de datos, siendo menos fiables para valores de concentración de NO_2 por encima de los 180 $\mu g/m^3$. Como trabajo futuro se proponen dos soluciones a este problema, la primera solución consiste en hacer un submuestreo de los datos correspondientes a concentraciones bajas y utilizar todos los datos que se corresponden a concentraciones altas. La segunda aproximación consiste en entrenar los modelos usando distintos pesos para los ejemplos, donde el peso de cada ejemplo es proporcional a la concentración de NO_2 indicada en la etiqueta del ejemplo. La segunda hipótesis se basa en atribuir las bajadas de concentración de NO_2 después de altas concentraciones a las medidas adoptadas por el Ayuntamiento de Madrid. De esta manera, si un día se alcanzan los 180 $\mu g/m^3$, el ayuntamiento pondrá en marcha las medidas de preaviso, reduciendo el tráfico y haciendo que se reduzca la concentración de NO_2 para el día siguiente. El modelo de predicción se ve afectado por estas acciones artificiales. Para entender el porqué se plantea el siguiente ejemplo: si el día 2 de febrero se alcanzase la cota de 180 $\mu g/m^3$, el Ayuntamiento pondría en marcha las medidas reduciendo la concentración de NO_2 para el día 3 de febrero. Sin embargo, el modelo al predecir la concentración de contaminante el 3 de febrero utilizaría los datos del día anterior, en los que el



(a) Gráfica de predicción contra realidad en la estación de calidad del aire Escuelas Aguirre.



(b) Realidad y la predicción de la concentración de NO_2 frente al tiempo en la estación de calidad del aire Escuelas Aguirre.

Figura 4.8: Valor real frente a la predicción de la concentración de NO_2 del modelo completo con ElasticNet de la concentración de NO_2.

tráfico era normal y la concentración de NO_2 alta, lo que haría predecir erróneamente al modelo que la concentración de NO_2 seguirá siendo alta el día 3 de febrero. Como trabajo futuro se podría resolver este problema utilizando una variable adicional que indique si el Ayuntamiento de Madrid ha puesto en marcha ciertas medidas.

4.4.2. Resultados Clasificación para la Clasificación de Alertas

Para intentar mejorar la predicción de alertas, el problema de la regresión se transforma en un problema de clasificación definiendo clases distintas en función de la concentración del NO_2. Para ello se definen las siguientes clases:

- Normal (N): Si la concentración de NO_2 está por debajo de los 180 $\mu g/m^3$.
- Alerta (A): En caso contrario.

Se elige el valor de 180 $\mu g/m^3$ porque es el límite para el preaviso según la normativa aprobada por el Ayuntamiento de Madrid [9]. Al hacer esta definición de las clases resulta que la estación Plaza del Carmen no tiene ninguna alerta, por lo que se omite del estudio de la clasificación. Los modelos utilizados son RandomForest y XGB explicados en la Sección 2.6. Como nos encontramos ante un conjunto con las clases muy desbalanceadas, ya que hay muy pocas alertas, si los clasificadores predijeran siempre clase Normal obtendrían un error muy pequeño, sin embargo un clasificador que prediga siempre lo mismo no tiene ningún interés. Para evitar que el desbalanceo en las clases enmascare el verdadero rendimiento de los clasificadores se calculan las matrices de confusión. Como puede observarse en las Tablas 4.6, 4.7 y 4.8 los resultados obtenidos por XGBoost son en general mejores que los obtenidos por el RandomForest. En ambos algoritmos se ajustan los parámetros de número de clasificadores base (árboles de decisión) y profundidad máxima de los árboles mediante el uso del algoritmo GridSearch. Los resultados que se exponen en este trabajo se obtienen utilizando 100 árboles con una profundidad máxima de 12 para el RandomForest, mientras que en XGB el resultado óptimo se obtiene con 300 árboles con profundidad máxima 5. Para entender este hecho primero hay que analizar qué está ocurriendo. El problema sistemático que se observa en las tablas es que la mayoría de los ejemplos se predicen como ejemplos normales y muy pocos como alertas. Un claro ejemplo de esta situación puede verse en la Tabla 4.8. El principal causante de este efecto radica en el desbalanceo de las clases, observamos por ejemplo que en la Tabla 4.7 hay 348 ejemplos normales y 19 de alertas. El método de XGB, al ser un algoritmo de boosting, se centra en acertar en los ejemplos que ha fallado inicialmente, lo que provoca que aunque los primeros clasificadores base estén sesgados y clasifiquen la mayoría de ejemplos como normales, según se realiza el boosting los ejemplos de alerta van adquiriendo un mayor peso en el entrenamiento, eliminándose el sesgo. Son notables los resultados obtenidos con XGB en la estación Escuelas Aguirre ya que, aunque se producen falsos positivos, todas las alertas se predicen correctamente.

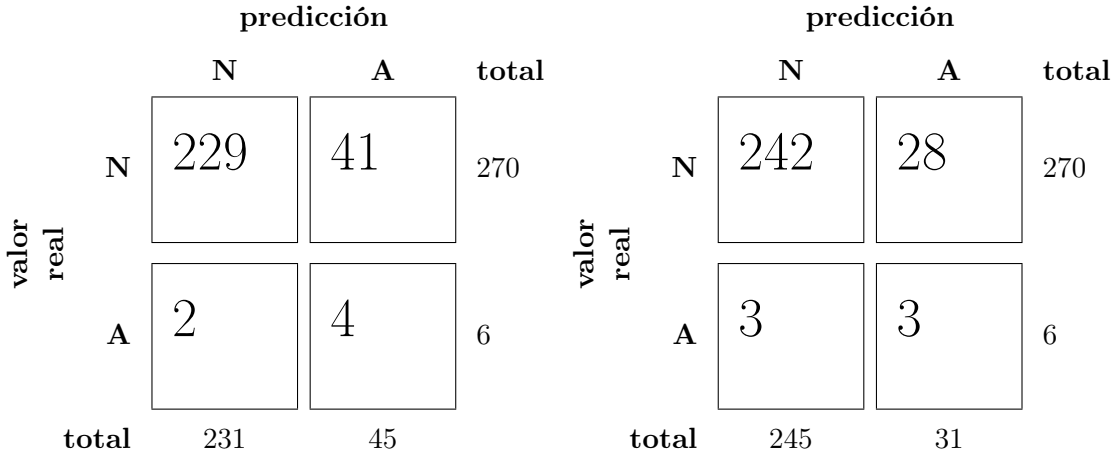


Tabla 4.6: Matrices de confusión para la clasificación de alertas usando RandomForest (izquierda) y XGB (derecha) en la estación Plaza de España

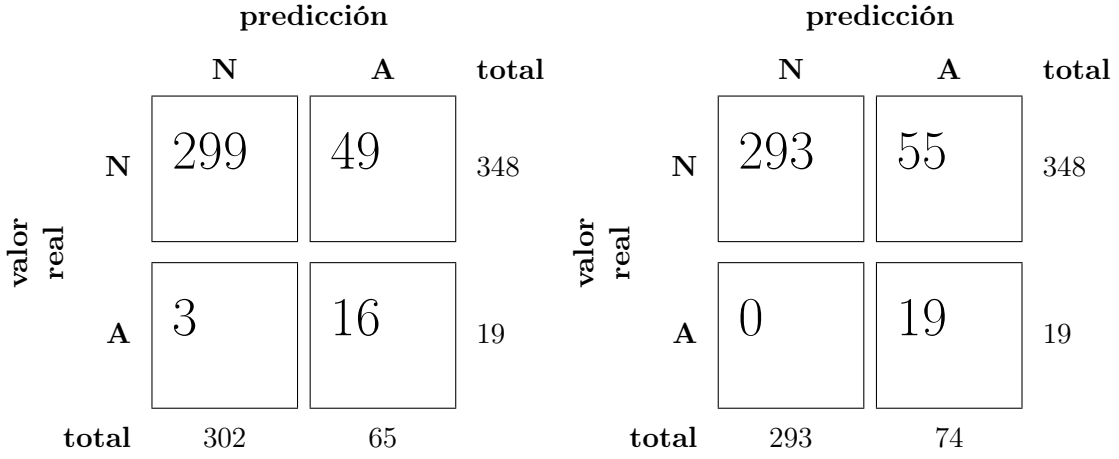


Tabla 4.7: Matrices de confusión para la clasificación de alertas usando RandomForest (izquierda) y XGB (derecha) en la estación Escuelas Aguirre

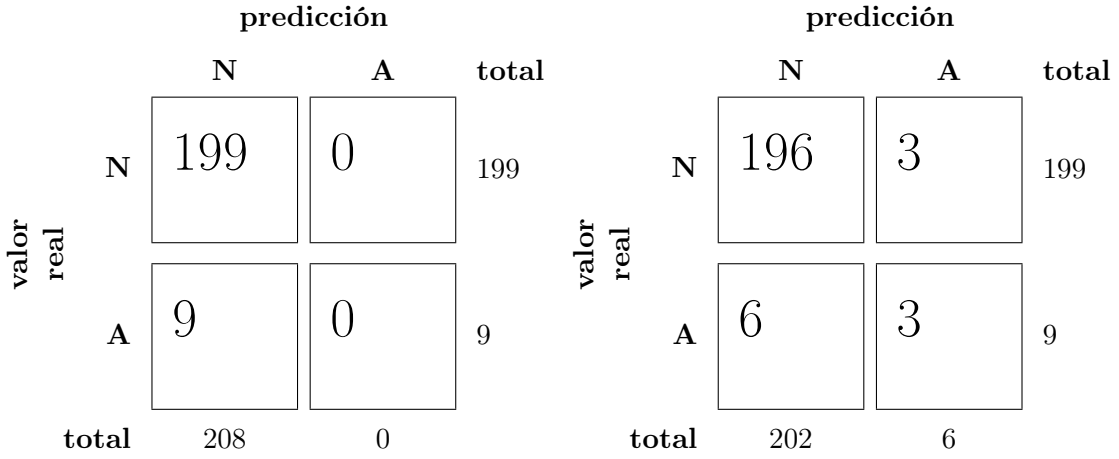


Tabla 4.8: Matrices de confusión para la clasificación de alertas usando RandomForest (izquierda) y XGB (derecha) en la estación Méndez Álvaro

5

Conclusiones y trabajo futuro

5.1. Conclusiones

El objetivo de este TFG tenía una triple vertiente, por una parte se pretende generar un modelo predictor con predicciones horarias de la concentración de NO_2 en Madrid. De manera similar, se pretende generar un modelo predictor de alertas horarias en Madrid. Por último pero no menos importante, se quiere determinar que variables de las utilizadas (carga de tráfico, temperatura, presión en la superficie, velocidad del viento y humedad relativa) son más relevantes a la hora de predecir la concentracion de NO_2. El desarrollo del TFG se ha realizado en tres etapas separadas. Una primera etapa de búsqueda de información en la que se revisó el estado del arte sobre predicción de contaminantes en ciudades, lo que sirvió para comprender el problema y buscar una solución adecuada. La segunda parte del desarrollo es la que conlleva mayor carga de trabajo y se centra en el desarrollo de la matriz de datos necesaria para aplicar las técnicas de aprendizaje automático. Fue necesario documentarse de los distintos formatos que usaba cada tipo de datos y transformarlos para poder combinarlos entre sí. Por último, la tercera fase se focaliza en el diseño experimental y la aplicación de los algoritmos de aprendizaje automático. Una vez que se tiene la matriz de datos, utilizando librerías de Python como scikit-learn, realizar los experimentos es más sencillo.

Los resultados obtenidos por tanto el modelo predictor de concentración de NO_2 como el modelo predictor de alertas ha sido positivo aunque altamente variables en función de la estación en la que se aplicasen. Estas variaciones en función de la estación se han analizado y tienen su explicación en la correlación de la concentración de NO_2 y la carga de tráfico en los puntos de medida cercanos a dicha estación. A mayor correlación se obtienen mejores resultados, tanto de regresión como de clasificación.

El modelo de clasificación intenta clasificar entre casos normales y alertas con éxito relativo. Se han utilizado los algoritmos RandomForest y XGB, obteniéndose los mejores resultados con XGB. Se cree que al ser las alertas una clase muy minoritaria, XGB es capaz de proporcionar mejor rendimiento puesto que se centra en los ejemplos más díficiles de clasificar. Cabe destacar el caso de Escuelas Aguirre donde XGB clasifica bien todas las alertas al precio de un número alto de falsos positivos.

En lo concerniente al modelo regresor, y también para evaluar la importancia de las variables,

se han comparado tres modelos distintos. Uno que sólo usa variables de la concentración del contaminante a pasado (Pasado), un segundo (Tráfico) que además que del contaminante a pasado incluye variables de carga de tráfico y un modelo completo (Completo) que incluye variables de concentración de contaminante a pasado, carga de tráfico y predicciones metereológicas. Además se prueban cuatro algoritmos de regresión distintos: Regresión Lineal, Lasso, ElasticNet y Regresión con Random Forest. Para comparar los resultados se utilizan dos medidas: el coeficiente de bondad de ajuste R^2 y el error relativo medio (MAPE), aunque ambas medidas miden cosas distintas los resultados entre ambas están fuertemente correlacionados. Se obtienen en media un valor R^2 de 0.55 y en el mejor caso por estaciones el MAPE alcanza un valor medio de 20 %. Los cuatro algoritmos obtienen un rendimiento muy similar, tanto que ElasticNet, a pesar de no obtener el mejor resultado en ninguna de las estaciones, es el más estable y el que mejores resultados obtiene en media. Un rasgo que se repite con todos los algoritmos es la existencia de un sesgo hacia valores bajos de concentración de NO_2, es decir, todos los algoritmos subestiman de forma sistemática las altas concentraciones de NO_2. Se ofrecen dos explicaciones a este hecho, la primera es la alta densidad de valores pequeños de concentración de NO_2, y la segunda tiene que ver con la influencia de las medidas llevadas a caba por el Ayuntamiento en la contaminación del día siguiente. Por otra parte, en los modelos de regresión, es destacable el hecho de que siempre se obtiene mejor resultado en el modelo completo excepto para la estación de Plaza de España, dónde el modelo Tráfico obtiene mejores rendimientos. Además, los modelos que incluyen las variables de tráfico siempre mejoran a los modelos basados únicamente en valores históricos de NO2. Este último resultado nos permite llegar al tercer objetivo del TFG, el análisis de las variables. No es evidente, a la luz de los resultados, que las variables metereológicas intervengan en la concentración de NO_2, sin embargo la carga de tráfico guarda una estrecha relación con esta concentración, hecho que se corrobora al estudiar las correlaciones entre ambas variables. Aquellas estaciones con mayor correlación entre la carga tráfico y la concentración de NO_2 obtiene mejores resultados, lo que indica que esta correlación en una estación nos permite hacernos una idea de la bondad del modelo en la estación y corrobora la relevancia de las variables de carga de tráfico.

Conviene mencionar los grandes problemas que se han encontrado a la hora de generar el modelo debido a la falta de homogeneidad en los datos. Hay muchos errores o huecos vacíos en los datos que hay que rellenar o descartar para poder construir el modelo. Esto provoca una construcción del modelo más compleja y más lenta. El desbalanceo de clases es otro factor con gran impacto en los resultados finales. Cuando se clasifican los datos en clases distintas según si suponen una alerta de contaminación o no, nos encontramos ante unos datos en los que la proporción de alertas suponen un $0,25\,\%$, lo que provoca que los modelos se centren únicamente en las mediciones de escenarios catalogados como normales, siendo muy dificil predecir las alertas. Ocurre lo mismo en la regresión, la mayoría de los ejemplos están en valores bajos de concentración del NO_2 por lo que al calcular regresiones que utilizan el criterio de mínimos cuadrados se minimiza el error en ejemplos de concentración baja, mientras que éste es grande en los ejemplos con concentraciones altas.

Durante el desarrollo de este Trabajo de Fin de Grado se han puesto en práctica muchos de los conocimientos obtenidos durante la carrera, especialmente los pertenecientes a la asignatura de 'Fundamentos de Aprendizaje Automático' donde, además de las bases teóricas y algoritmos básicos del aprendizaje automático, también se aprende a utilizar Python y librerías especializadas como scikit-learn o numpy. Han sido fundamentales también los conocimientos de las asignaturas de Estadística I y Estadística II, donde se profundiza en la teoría de la regresión. También ha sido necesario aprender a usar nuevas herramientas como la librería Pandas, que ha jugado un papel muy relevante durante todo el desarrollo, especialmente en la fase de transformación de los datos.

5.2. Retos futuros

Queda como trabajo futuro hacer algunos cambios en el modelo con la idea de eliminar el sesgo del modelo a concentraciones bajas, así como incluir la información sobre las medidas tomadas por el Ayuntamiento. Adicionalmente, sería útil utilizar un modelo predictor de la carga de tráfico en lugar de usar el tráfico del día anterior, asunto sobre el que se discute en el Anexo C.

Más aún, es interesante extender el sistema de predicción a todas las estaciones de contaminación, así como el de ampliarlo a la predicción de otros contaminantes como el CO_2. También se puede hacer un sistema en tiempo real que sería muy útil, aunque limitado por los tiempos de refresco de la información. Por último, es importante realizar un estudio de qué estaciones de tráfico son más útiles para predecir el tráfico en cada estación de medición de la calidad del aire, ya que las más cercanas pueden no resolver este problema.

Glosario de acrónimos

- **NOAA**: National Oceanic and Atmospheric Administration
- **GFS**: Global Forecast System
- **TFG**: Trabajo de Fin de Grado

Bibliografía

[1] Portal de datos abiertos. http://datos.madrid.es/FWProjects/egob/contenidos/datasets/ficheros/Transporte_trafico/PUNTOS%20MEDIDA%20TRAFICO_MADRID.pdf, 2012.

[2] Portal de datos abiertos. http://datos.madrid.es/FWProjects/egob/contenidos/datasets/ficheros/Estructura%20y%20contenido%20del%20fichero%20csv.pdf, 2012.

[3] Portal de datos abiertos. http://datos.madrid.es/FWProjects/egob/contenidos/datasets/ficheros/Interprete_ficheros_%20calidad_%20del_%20aire_global.pdf, 2012.

[4] ELENA G. SEVILLANO. El mapa de la contaminación en europa. http://internacional.elpais.com/internacional/2017/01/13/actualidad/1484338094_275966.html, 2017. Accedido: 18-02-2017.

[5] Margaret Bell, Angela S Bergantino, Mario Catalano, Fabio Galatioto, et al. Prediction of air pollution peaks generated by urban transport networks. Technical report, 2015.

[6] Ole Raaschou-Nielsen, Zorana J Andersen, Rob Beelen, Evangelia Samoli, Massimo Stafoggia, Gudrun Weinmayr, Barbara Hoffmann, Paul Fischer, Mark J Nieuwenhuijsen, Bert Brunekreef, et al. Air pollution and lung cancer incidence in 17 european cohorts: prospective analyses from the european study of cohorts for air pollution effects (escape). *The lancet oncology*, 14(9):813–822, 2013.

[7] Rob Beelen, Ole Raaschou-Nielsen, Massimo Stafoggia, Zorana Jovanovic Andersen, Gudrun Weinmayr, Barbara Hoffmann, Kathrin Wolf, Evangelia Samoli, Paul Fischer, Mark Nieuwenhuijsen, et al. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project. *The Lancet*, 383(9919):785–795, 2014.

[8] Feng Lu, Dongqun Xu, Yibin Cheng, Shaoxia Dong, Chao Guo, Xue Jiang, and Xiaoying Zheng. Systematic review and meta-analysis of the adverse health effects of ambient pm 2.5 and pm 10 pollution in the chinese population. *Environmental research*, 136:196–204, 2015.

[9] Dirección General de Sostenibilidad. Área de Gobierno de Medio Ambiente y Movilidad del Ayuntamiento de Madrid. EvaluaciÓn del impacto en los niveles de concentraciÓn de no 2. http://www.madrid.es/UnidadesDescentralizadas/UDCMedios/noticias/2016/01Enero/21Jueves/Notasprensa/Nuevo%20protocolo%20contaminaci%C3%B3n/ficheros/Nuevo%20protocolo%20NO2.pdf, 2017. Accedido 21-04-2017.

[10] Pancrazio Bertaccini, Vanja Dukic, and Rosaria Ignaccolo. Modeling the short-term effect of traffic and meteorology on air pollution in turin with generalized additive models. *Advances in Meteorology*, 2012, 2012.

[11] Portal de datos abiertos. http://datos.madrid.es/portal/site/egob, 2012.

[12] Saleh M Al-Alawi, Sabah A Abdul-Wahab, and Charles S Bakheit. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, 23(4):396–403, 2008.

[13] Dan Wei. Predicting air pollution level in a specific city.

[14] ÁLVARO SÁNCHEZ Twitter. Bruselas da un ultimátum a españa para que mejore la calidad del aire. `http://politica.elpais.com/politica/2017/02/15/actualidad/1487169552_832008.html`, 2016. Accedido: 18-02-2017.

[15] István Juhos, László Makra, and Balázs Tóth. Forecasting of traffic origin no and no 2 concentrations by support vector machines and neural networks using principal component analysis. *Simulation Modelling Practice and Theory*, 16(9):1488–1502, 2008.

[16] Kunwar P Singh, Shikha Gupta, and Premanjali Rai. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, 80:426–437, 2013.

[17] SIV Sousa, FG Martins, MC Pereira, and MCM Alvim-Ferraz. Prediction of ozone concentrations in oporto city with statistical approaches. *Chemosphere*, 64(7):1141–1149, 2006.

[18] World Health Organization et al. Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide: report on a who working group, bonn, germany 13-15 january 2003. 2003.

[19] Noaa gfs. `https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs`, 2013.

[20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[21] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[22] Leo Breiman. Classification and regression trees. 1984.

[23] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

[24] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[25] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed <today>].

[26] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[27] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. van der Voort S, Millman J, 2010.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[29] John D Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

A

Ejemplos de matrices de datos procesadas

A.1. Datos de tráfico

Un ejemplo de los datos del tráfico procesados es el siguiente se puede observar en la tabla A.1.

A.2. Datos de calidad del aire

Un ejemplo de los datos de calidad del aire procesados puede encontrarse en la tabla A.3.

idelem	fecha	carga	carga_n1s1	carga_n2s1	carga_n3s1	carga_n6s1	carga_n1s24
3395	2016-10-01 00:00:00	20.0	nan	nan	nan	nan	nan
3395	2016-10-01 01:00:00	18.0	20.0	nan	nan	nan	nan
3395	2016-10-01 02:00:00	14.0	18.0	20.0	nan	nan	nan
3395	2016-10-01 03:00:00	10.0	14.0	18.0	20.0	nan	nan
3395	2016-10-01 04:00:00	10.0	10.0	14.0	18.0	nan	nan
3395	2016-10-01 05:00:00	8.0	10.0	10.0	14.0	nan	nan
3395	2016-10-01 06:00:00	4.0	8.0	10.0	10.0	20.0	nan
3395	2016-10-01 07:00:00	7.0	4.0	8.0	10.0	18.0	nan
3395	2016-10-01 08:00:00	7.0	7.0	4.0	8.0	14.0	nan
3395	2016-10-01 09:00:00	9.0	7.0	7.0	4.0	10.0	nan
3395	2016-10-01 10:00:00	15.0	9.0	7.0	7.0	10.0	nan
3395	2016-10-01 11:00:00	21.0	15.0	9.0	7.0	8.0	nan
3395	2016-10-01 12:00:00	30.0	21.0	15.0	9.0	4.0	nan
3395	2016-10-01 13:00:00	42.0	30.0	21.0	15.0	7.0	nan
3395	2016-10-01 14:00:00	41.0	42.0	30.0	21.0	7.0	nan
3395	2016-10-01 15:00:00	30.0	41.0	42.0	30.0	9.0	nan
3395	2016-10-01 16:00:00	15.0	30.0	41.0	42.0	15.0	nan
3395	2016-10-01 17:00:00	19.0	15.0	30.0	41.0	21.0	nan
3395	2016-10-01 18:00:00	19.0	19.0	15.0	30.0	30.0	nan
3395	2016-10-01 19:00:00	27.0	19.0	19.0	15.0	42.0	nan
3395	2016-10-01 20:00:00	27.0	27.0	19.0	19.0	41.0	nan
3395	2016-10-01 21:00:00	38.0	27.0	27.0	19.0	30.0	nan
3395	2016-10-01 22:00:00	39.0	38.0	27.0	27.0	15.0	nan
3395	2016-10-01 23:00:00	27.0	39.0	38.0	27.0	19.0	nan
3395	2016-10-02 00:00:00	17.0	27.0	39.0	38.0	19.0	20.0

Tabla A.1: Ejemplo de matriz de datos de tráfico procesados

Tabla A.2: My caption

station_code	date	hour_val	hour_val_n0s1	hour_val_n1s1	hour_val_n2s1	hour_val_n3s1	hour_val_n4s1	hour_val_n1s24	hour_val_n2s24
28079004	2016-11-01 00:00:00	77.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 01:00:00	65.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 02:00:00	63.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 03:00:00	54.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 04:00:00	48.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 05:00:00	41.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 06:00:00	37.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 07:00:00	42.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 08:00:00	37.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 09:00:00	41.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 10:00:00	44.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 11:00:00	46.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 12:00:00	49.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 13:00:00	45.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 14:00:00	46.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 15:00:00	41.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 16:00:00	45.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 17:00:00	53.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 18:00:00	79.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 19:00:00	103.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 20:00:00	106.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 21:00:00	94.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 22:00:00	72.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-01 23:00:00	65.0	nan	nan	nan	nan	nan	nan	nan
28079004	2016-11-02 00:00:00	59.0	77.0	nan	nan	nan	nan	nan	nan
28079004	2016-11-02 01:00:00	50.0	65.0	77.0	nan	nan	nan	nan	nan
28079004	2016-11-02 02:00:00	42.0	63.0	65.0	77.0	nan	nan	nan	nan
28079004	2016-11-02 03:00:00	34.0	54.0	63.0	65.0	77.0	nan	nan	nan
28079004	2016-11-02 04:00:00	30.0	48.0	54.0	63.0	65.0	77.0	nan	nan
28079004	2016-11-02 05:00:00	25.0	41.0	48.0	54.0	63.0	65.0	nan	nan
28079004	2016-11-02 06:00:00	29.0	37.0	41.0	48.0	54.0	63.0	nan	nan
28079004	2016-11-02 07:00:00	41.0	42.0	37.0	41.0	48.0	54.0	nan	nan
28079004	2016-11-02 08:00:00	59.0	37.0	42.0	37.0	41.0	48.0	nan	nan
28079004	2016-11-02 09:00:00	65.0	41.0	37.0	42.0	37.0	41.0	nan	nan
28079004	2016-11-02 10:00:00	80.0	44.0	41.0	37.0	42.0	37.0	nan	nan
28079004	2016-11-02 11:00:00	87.0	46.0	44.0	41.0	37.0	42.0	nan	nan
28079004	2016-11-02 12:00:00	90.0	49.0	46.0	44.0	41.0	37.0	nan	nan
28079004	2016-11-02 13:00:00	74.0	45.0	49.0	46.0	44.0	41.0	nan	nan
28079004	2016-11-02 14:00:00	61.0	46.0	45.0	49.0	46.0	44.0	nan	nan
28079004	2016-11-02 15:00:00	59.0	41.0	46.0	45.0	49.0	46.0	nan	nan
28079004	2016-11-02 16:00:00	53.0	45.0	41.0	46.0	45.0	49.0	nan	nan
28079004	2016-11-02 17:00:00	64.0	53.0	45.0	41.0	46.0	45.0	nan	nan
28079004	2016-11-02 18:00:00	90.0	79.0	53.0	45.0	41.0	46.0	nan	nan
28079004	2016-11-02 19:00:00	97.0	103.0	79.0	53.0	45.0	41.0	nan	nan
28079004	2016-11-02 20:00:00	107.0	106.0	103.0	79.0	53.0	45.0	nan	nan
28079004	2016-11-02 21:00:00	108.0	94.0	106.0	103.0	79.0	53.0	nan	nan
28079004	2016-11-02 22:00:00	101.0	72.0	94.0	106.0	103.0	79.0	nan	nan
28079004	2016-11-02 23:00:00	88.0	65.0	72.0	94.0	106.0	103.0	nan	nan
28079004	2016-11-03 00:00:00	61.0	59.0	65.0	72.0	94.0	106.0	77.0	nan

Tabla A.3: Ejemplo de matriz de datos de calidad del aire procesados

B

Cálculo de distancias

Cálculo de distancias

Para el cálculo de distancias ha sido falta convertir las coordenadas, las estaciones de calidad del aire tienen sus coordenadas en representación DMS (Degree Minutes Seconds), mientras que las estaciones de tráfico tienen sus coordenadas en representación UTM. En este trabajo, ambas representaciones se convierten a latitud y longitud para después calcular su distancia siguiendo el algoritmo de Vincenty.

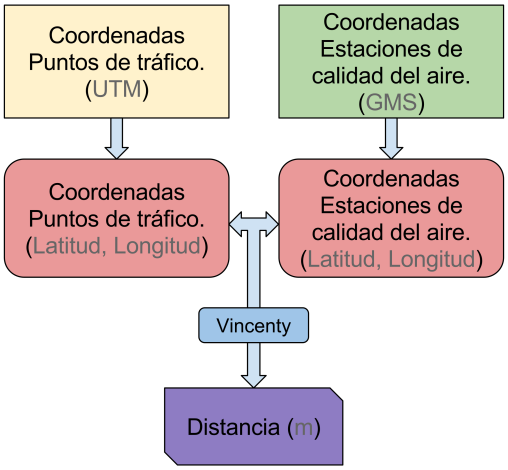


Figura B.1: Esquema cálculo de distancias

Hay que expresar tanto las coordenadas de puntos de tráfico como las de estaciones de calidad del aire en latitudes y longitudes antes de poder calcular las distancias.

C

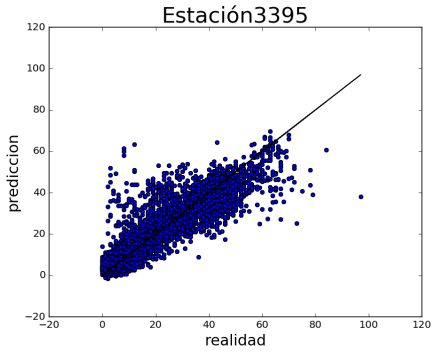
Modelo del tráfico

C.1. Introducción

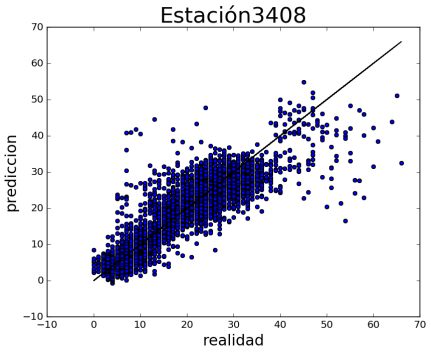
Como la solución de utilizar el tráfico 24 horas antes no parecía óptima, se genera un modelo predictor para la carga del tráfico. Las predicciones de este modelo se utilizan después para la predicción de la concentracion de NO_2.

Para predecir el tráfico, el score R^2 medio obtenido por el modelo de tráfico ha sido 0.760305435367 mientras que el obtenido al utilizar el tráfico 24 horas antes ha sido 0.101577534224. Esto se puede apreciar en las Figuras C.1 y C.3.

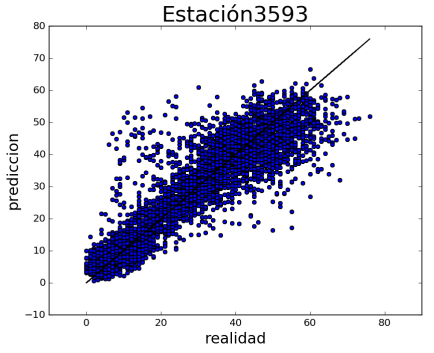
Se han realizado las pruebas con ElasticNet y el modelo completo. Prediciendo el tráfico desde el 01/02/2016 hasta el 31/12/2016. El modelo de predicción de concentración de NO_2 utiliza como entrenamiento desde el 01/02/2016 hasta 30/9/2016 y el resto como validación. Las gráficas C.3a, C.3b, C.3c y C.3d muestran los resultados obtenidos. En esta primer aproximación los resultados son similares al utilizar el modelo del tráfico en lugar de los datos de carga de tráfico a pasado.



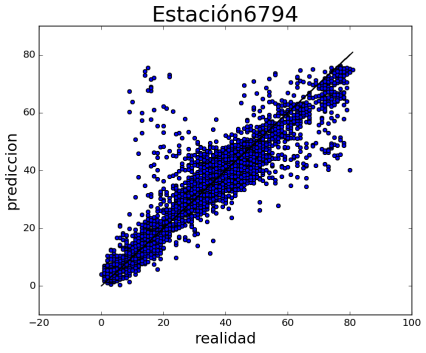
(a) Predicción contra realidad en el modelo de tráfico en la estación con identificador 3395.



(b) Predicción contra realidad en el modelo de tráfico en la estación con identificador 3408.

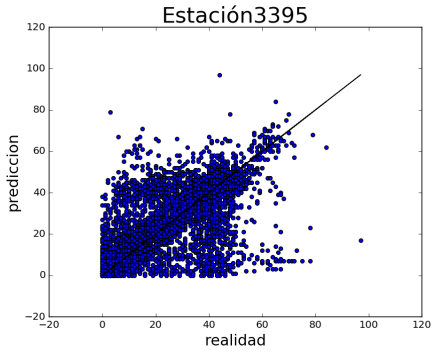


(c) Predicción contra realidad en el modelo de tráfico en la estación con identificador 3593.

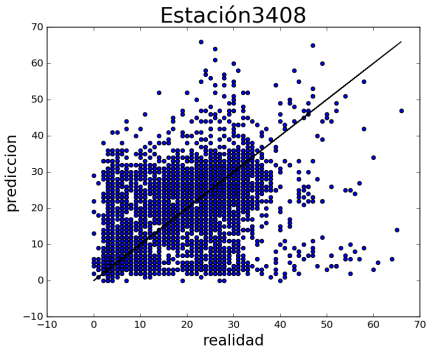


(d) Predicción contra realidad en el modelo de tráfico en la estación con identificador 6794.

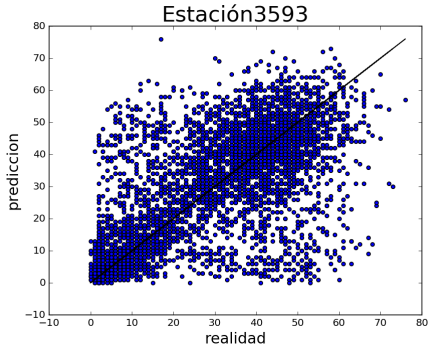
Figura C.1: Predicción contra realidad en el modelo de tráfico.



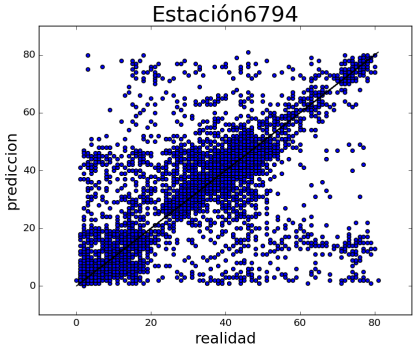
(a) Predicción contra realidad en la estación con identificador 3395 utilizando el trafico 24 horas antes.



(b) Predicción contra realida en la estación con identificador 3408 utilizando el trafico 24 horas antes.

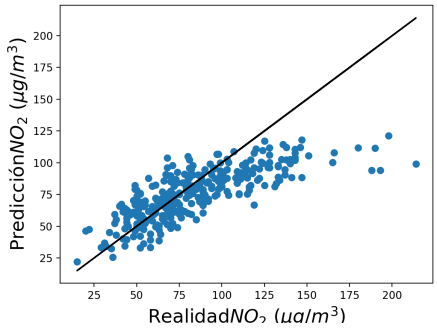


(c) Predicción contra realidad en la estación con identificador 3593 utilizando el trafico 24 horas antes.

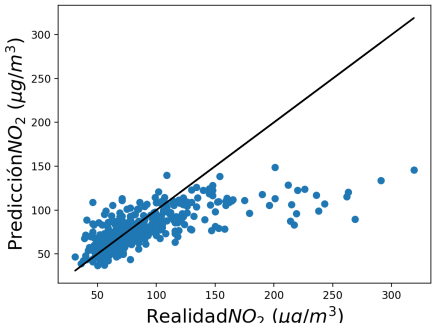


(d) Predicción contra realidad en la estación con identificador 3395 utilizando el trafico 24 horas antes.

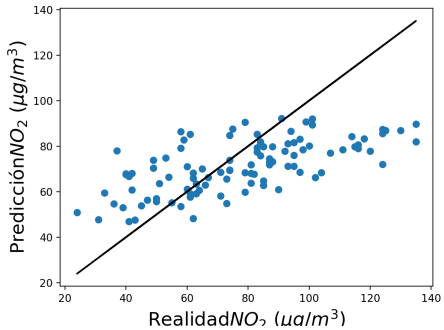
Figura C.2: Predicción contra realidad en el modelo de tráfico.



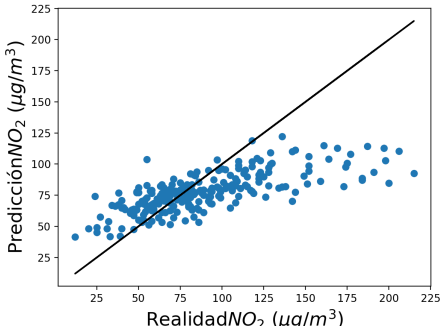
(a) Predicción contra realidad en la estación de calidad del aire de Plaza de España. Los resultados son $R^2 = 0.5138$ y MAPE=20.28 %.



(b) Predicción contra realidad en la estación de calidad del aire de Escuelas Aguirre. Los resultados son $R^2 = 0.35$ y MAPE=21.22 %



(c) Predicción contra realidad en la estación de calidad del aire de Plaza del Carmen. Los resultados son $R^2 = 0.3287$ y MAPE=17.08 %



(d) Predicción contra realidad en la estación de calidad del aire de Méndez Álvaro. Los resultados son $R^2 = 0.3482$ y MAPE=26.98 %

Figura C.3: Predicción contra realidad en el modelo de predicción de NO_2 completo utilizando el modelo de predicción de tráfico para obtener la carga de tráfico.

D

Medidas frente a la concentración de NO_2

Extraído de [9]

D.1. INTRODUCCIÓN

La Ley 34/2007, de 15 de noviembre, de Calidad del Aire y Protección de la Atmósfera, que tiene como uno de sus principios rectores el de cautela y acción preventiva, establece, en el ámbito de la Administración local, para los municipios de más de 100.000 habitantes y las aglomeraciones, determinadas obligaciones como las de disponer de instalaciones y redes de evaluación, informar a la población sobre los niveles de contaminación y calidad del aire, elaborar planes y programas para los objetivos de calidad del aire, e integrar las consideraciones relativas a la protección atmosférica en la planificación de las distintas políticas sectoriales, adoptando cuando sea necesario medidas de restricción total o parcial del tráfico. De igual modo, la Ley 6/2014 por la que se modifica el texto articulado de la Ley sobre Tráfico, Circulación de Vehículos a Motor y Seguridad Vial, aprobado por el Real Decreto Legislativo 339/1990, de 2 de marzo, atribuye a los municipios la competencia de restricción de la circulación a determinados vehículos en vías urbanas por motivos medioambientales.

El Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire, establece umbrales de alerta para tres contaminantes, dióxido de nitrógeno, dióxido de azufre y ozono, y define el umbral de alerta como el nivel a partir del cual una exposición de breve duración supone un riesgo para la salud humana, que afecta al conjunto de la población y que requiere la adopción de medidas inmediatas. El valor del umbral de alerta para el dióxido de nitrógeno está establecido en 400 microgramos/m3 durante tres horas consecutivas en lugares representativos de la calidad del aire, en un área de al menos 100 km2 o en una zona o aglomeración entera, si esta última superficie es menor. El citado Real Decreto establece asimismo un valor límite horario para la protección de la salud de dióxido de nitrógeno de 200 microgramos/m3 que no debe superarse más de 18 horas al año en ninguna de las estaciones de la red.

El Ayuntamiento de Madrid, para llevar a cabo el control de la calidad del aire de la ciudad, dispone del Sistema de Vigilancia, Predicción e Información de la Calidad del Aire que permite conocer, de forma continua y en tiempo real, las concentraciones de contaminantes, con el principal objetivo de proteger la salud de la población y reducir al máximo las situaciones de riesgo.

El umbral de alerta para el dióxido de nitrógeno no se ha superado en ninguna ocasión en el municipio de Madrid pero sí el valor límite horario en varias de las estaciones de la red. Las elevadas concentraciones son debidas fundamentalmente a las emisiones del tráfico, y 2 tienen lugar en situaciones con condiciones meteorológicas especialmente adversas, que requieren la ejecución de medidas para reducir los niveles de contaminación y la duración de los episodios, y evitar que llegue a superarse el valor límite horario y que se llegue a alcanzar el umbral de alerta. Para ello, se establece una división en zonas del territorio municipal de tal manera que las situaciones de alerta puedan declararse en áreas más reducidas con alta densidad de población. Igualmente se definen unos niveles de aviso que permitan, en el caso de registrarse concentraciones elevadas de dióxido de nitrógeno, la puesta en marcha de mecanismos de información adicionales, que sirvan tanto para proteger la salud de los ciudadanos como para sensibilizar a la opinión pública, recabar su colaboración para la reducción de la contaminación y, en función de los niveles alcanzados y la duración del episodio, llevar a cabo medidas de restricción de tráfico en la ciudad y sus accesos para reducir los niveles de contaminación y evitar que se alcance la situación de alerta.

D.2. ZONIFICACIÓN DE LA CIUDAD DE MADRID

La ciudad de Madrid, a los efectos de este Protocolo, se ha dividido en cinco zonas teniendo en consideración: - La distribución de la población

- La tipología y distribución de estaciones del sistema de vigilancia de la calidad del aire

- El viario de tráfico, para facilitar la implantación de posibles actuaciones de restricción del mismo

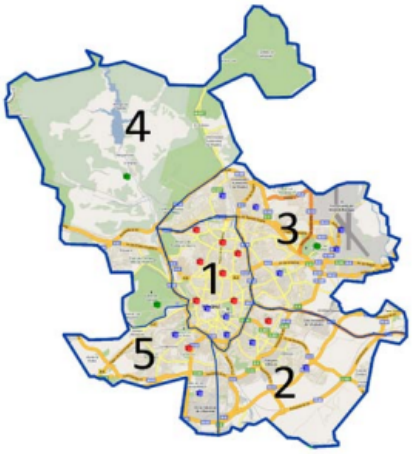


Figura D.1: Zonificación de Madrid

D.3. DEFINICIÓN DE NIVELES DE ACTUACIÓN

Se establecen tres niveles de actuación en función de las concentraciones de dióxido de nitrógeno que se registren en las zonas que se han definido. **NIVELES:** PREAVISO: cuando en dos estaciones cualesquiera de una misma zona se superan los 180 microgramos/m3 durante dos horas consecutivas.

AVISO: cuando en dos estaciones cualesquiera de una misma zona se superan los 200 microgramos/m3 durante dos horas consecutivas.

ALERTA: cuando en tres estaciones cualesquiera de una misma zona (o dos si se trata de la zona 4) se superan los 400 microgramos/m3 durante tres horas consecutivas.

D.4. ESCENARIOS POSIBLES

Una vez superado alguno de los niveles citados, y si la previsión meteorológica es desfavorable 1, se considerará iniciado un episodio de contaminación. Para la puesta en marcha de las actuaciones que para cada uno de los escenarios a continuación se detallan, se tienen en cuenta los valores alcanzados así como la persistencia de las superaciones.

ESCENARIO 1: 1 día con superación del nivel de preaviso Actuaciones: - Medidas Informativas 1

- Reducción de la velocidad a 70 km/h en la M?30 y accesos

- Medidas de Promoción del Transporte Público.

ESCENARIO 2: 2 días consecutivos con superación del nivel de preaviso ó 1 día con superación del nivel de aviso Actuaciones: - Medidas Informativas 1 y 2

- Reducción de la velocidad a 70 km/h en la M?30 y accesos

- Prohibición del estacionamiento de vehículos en las plazas y horario del Servicio de Estacionamiento Regulado (SER) en el interior de la M?30

- Medidas de Promoción del Transporte Público.

ESCENARIO 3: 2 días consecutivos con superación del nivel de aviso Actuaciones: - Medidas Informativas 1 y 2

- Reducción de la velocidad a 70 km/h en la M?30 y accesos

- Prohibición del estacionamiento de vehículos en las plazas y horario del SER en el interior de la M?30

- Restricción de la circulación en el interior de la almendra central (área interior de la M?30) del 50 % de todos los vehículos

- Medidas de Promoción del Transporte Público

- Se recomienda la no circulación de taxis libres, excepto Ecotaxis y Eurotaxis, en el interior de la almendra central (área interior de la M?30), pudiendo estos vehículos estacionar en las plazas azules del SER, además de en sus paradas habituales, a la espera de viajeros.

ESCENARIO 4: 3 días consecutivos de nivel de aviso o 1 día de nivel de alerta Actuaciones: - Medidas Informativas 1 y 2

- Reducción de la velocidad a 70 km/h en la M?30 y accesos

- Prohibición del estacionamiento de vehículos en las plazas y horario del SER en el interior de la M?30

- Restricción de la circulación en el interior de la almendra central (área interior de la M?30) del 50 % de todos los vehículos

- Restricción de la circulación por la M?30 del 50 % de todos los vehículos

- Restricción de la circulación de taxis libres, excepto Ecotaxis y Eurotaxis, en el interior de la almendra central (área interior de la M?30)

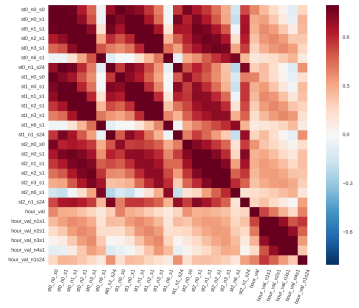
- Medidas de Promoción del Transporte Público

E

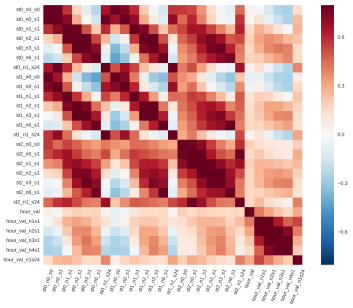
Correlación entre variables

Los modelos lineales como la regresión lineal o el método Lasso explicados en la Sección 2.7 utilizan la correlación entre las variables regresoras y la variable explicada para calcular los coeficientes que minimizan el error de regresión. Además, la correlación entre las variables y la variable explicada expresa la dependencia lineal entre ambas y las correlaciones entre variables regresoras representan dependencias lineales. Para conocer que variables aportan más información sobre la concentración de NO_2 y para saber que variables están correlacionadas entre sí, se calcula la matriz de correlación y se representa como un mapa de calor. La Figura E.1 muestra la correlacion entre la carga del tráfico y la concentración del contaminante, mientras que la Figura E.2 muestra la correlación entre las variables meterológicas y esta misma concentración. Los mapas de calor se segregan por estaciones de calidad del aire de esta manera se pueden relacionar con los resultados expuestos en la Sección 4.4. En la Figura E.1 se puede observar que la correlación entre el contaminante

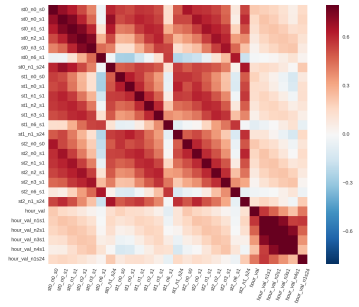
Correlaciones lineales



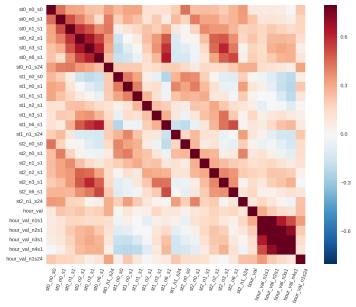
(a) Mapa de calor de la correlación entre concentración de NO_2 y la carga del tráfico en la estación con identificador 28079008.



(b) Mapa de calor de la correlación entre concentración de NO_2 y la carga del tráfico en la estación con identificador 28079008 para valores de concentración de NO_2 mayores que 100 $\mu g/m^3$.

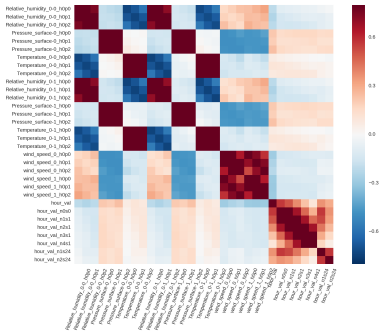


(c) Mapa de calor de la correlación entre concentración de NO_2 y la carga del tráfico en la estación con identificador 28079035.

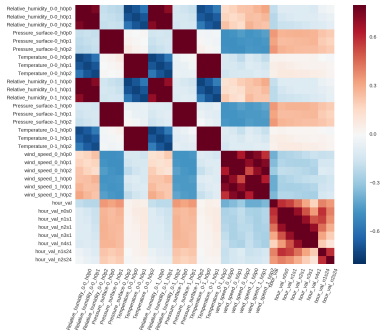


(d) Mapa de calor de la correlación entre concentración de NO_2 y la carga del tráfico en la estación con identificador 28079035 para valores de concentración de NO_2 mayores que 100 $\mu g/m^3$

Figura E.1: Mapas de calor de correlación entre la carga de tráfico y la concentración de NO_2 en las estaciones 28079008 y 28079035.



(a) Mapa de calor de la correlación entre concentración de NO_2 y las variables metereológicas en la estación con identificador 28079008.



(b) Mapa de calor de la correlación entre concentración de NO_2 y las variables metereológicas en la estación con identificador 28079035.

Figura E.2: Mapas de calor de correlación entre variables metereo?ogicas y la concentración de NO_2 en las estaciones 28079008 y 28079035.

E.0.1. Correlación entre carga de tráfico y concentración de NO_2

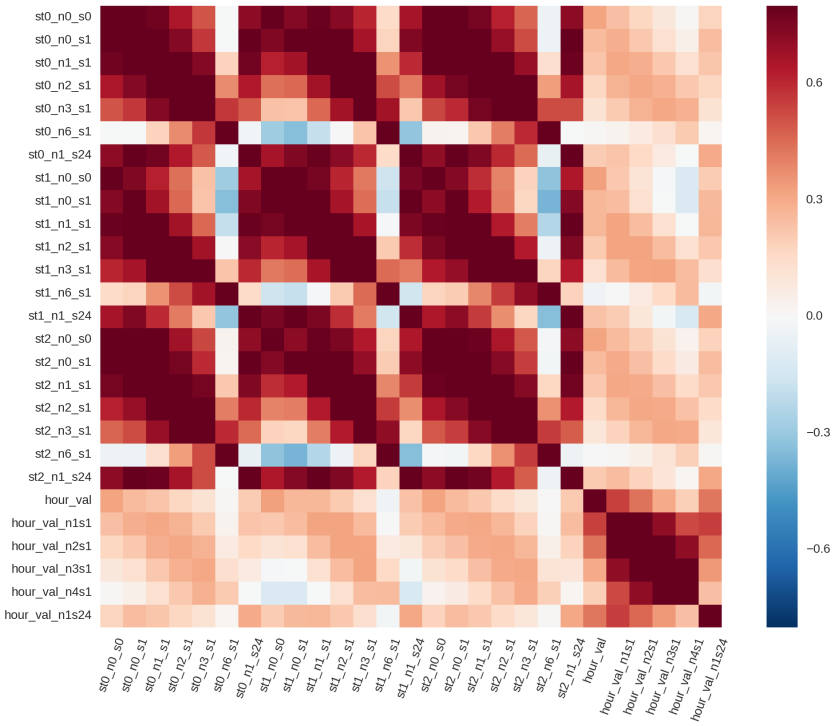


Figura E.3: Correlación entre carga de tráfico y contaminante en la estación 28079004.

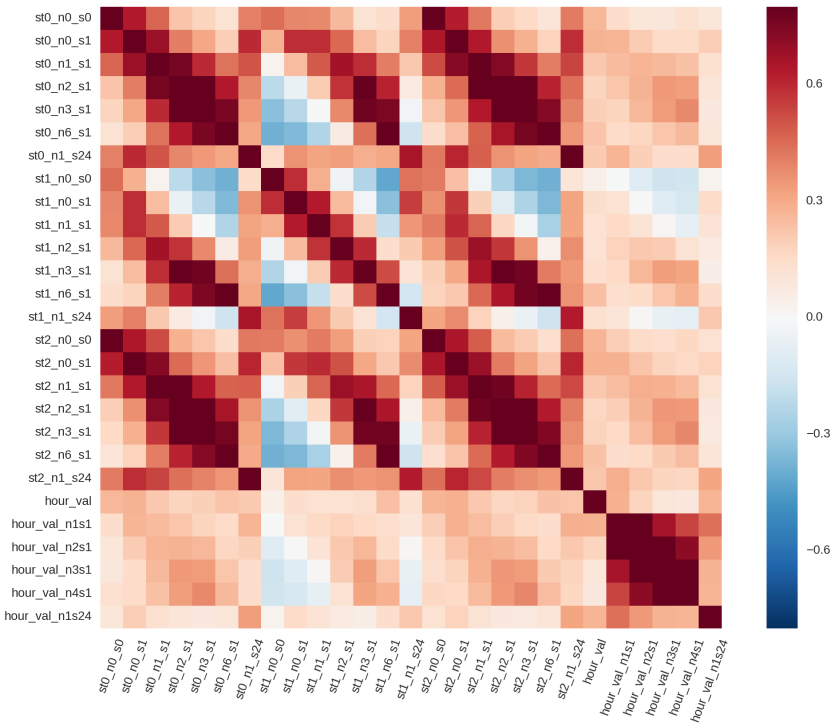


Figura E.4: Correlación entre carga de tráfico y contaminante en la estación 28079004 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.

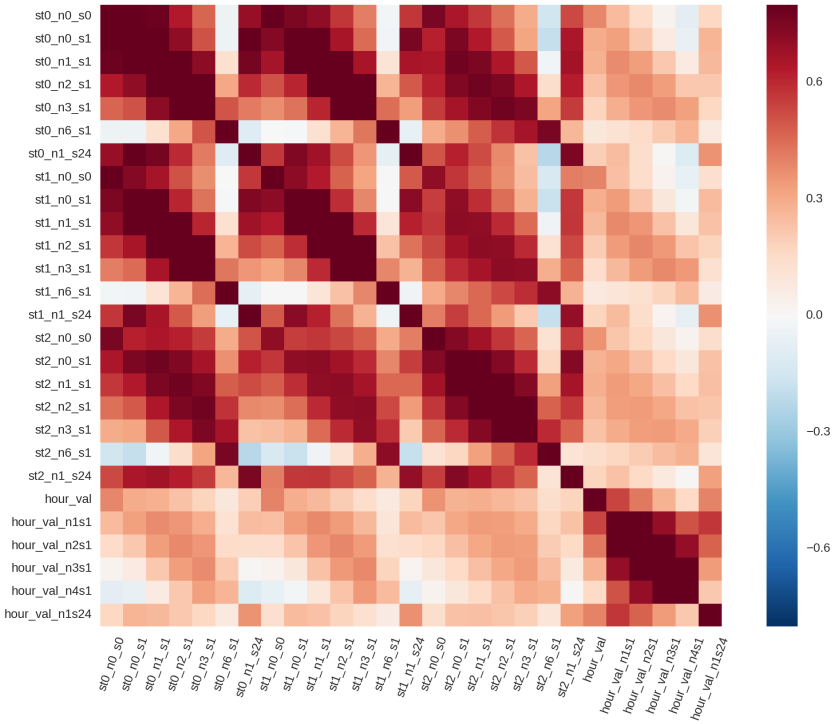


Figura E.5: Correlación entre carga de tráfico y contaminante en la estación 28079008.

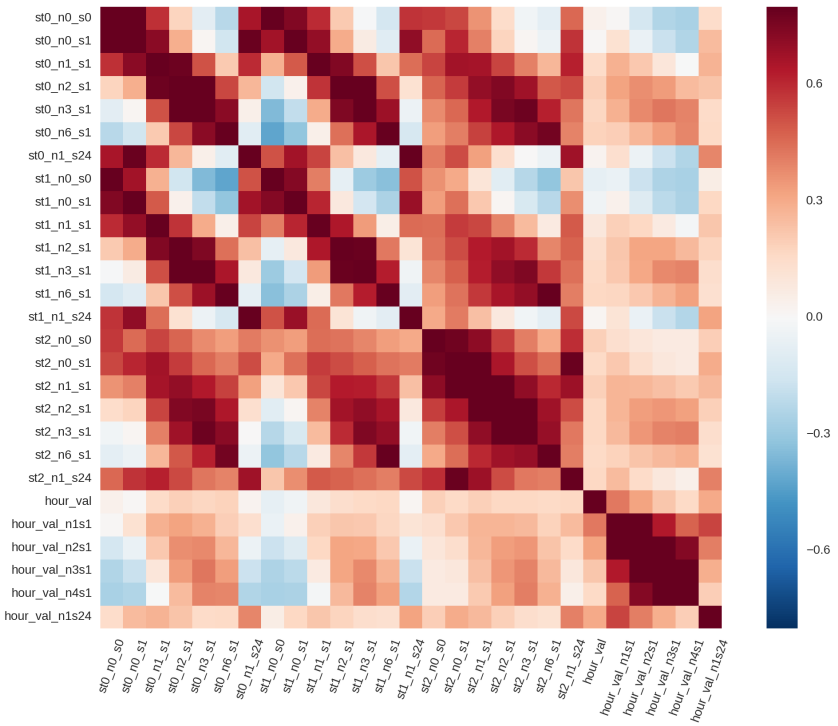


Figura E.6: Correlación entre carga de tráfico y contaminante en la estación 28079008 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.

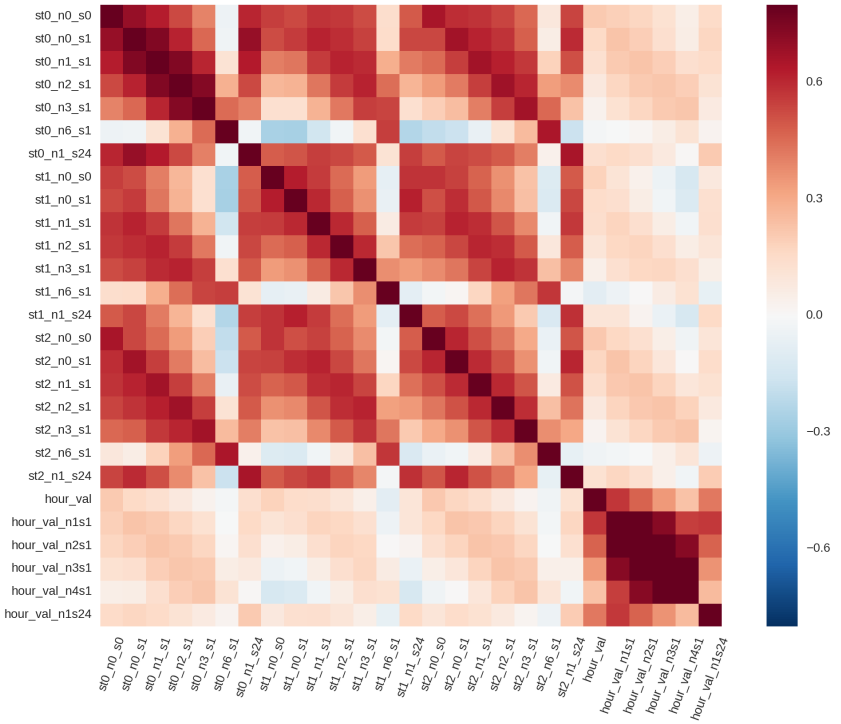


Figura E.7: Correlación entre carga de tráfico y contaminante en la estación 28079035.

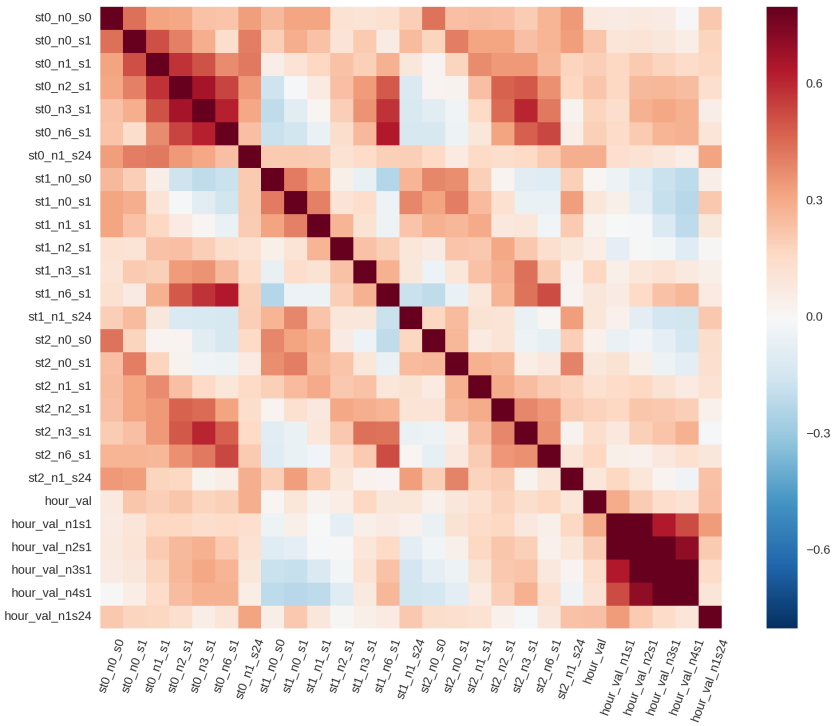


Figura E.8: Correlación entre carga de tráfico y contaminante en la estación 28079035 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.

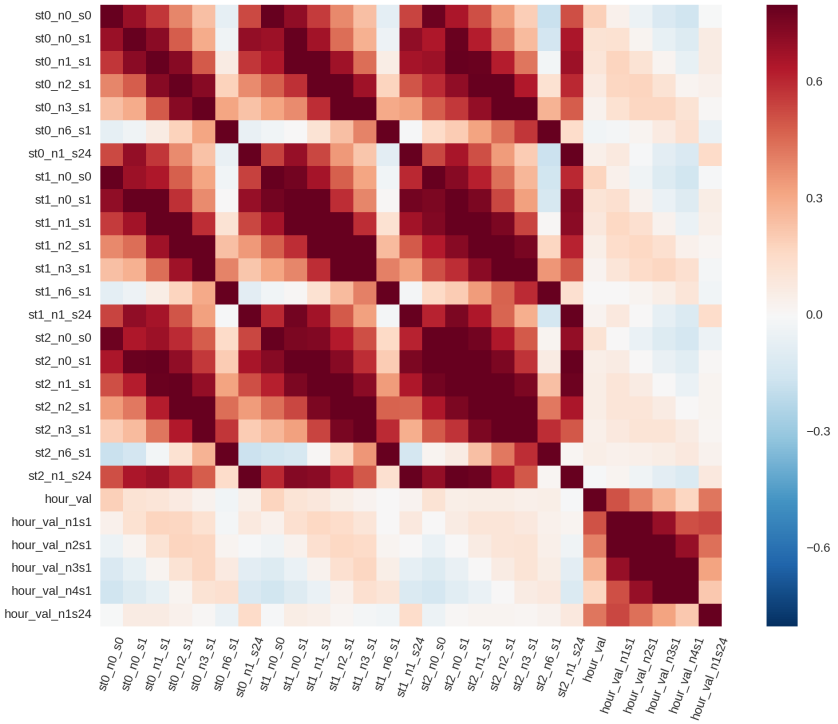


Figura E.9: Correlación entre carga de tráfico y contaminante en la estación 28079047.

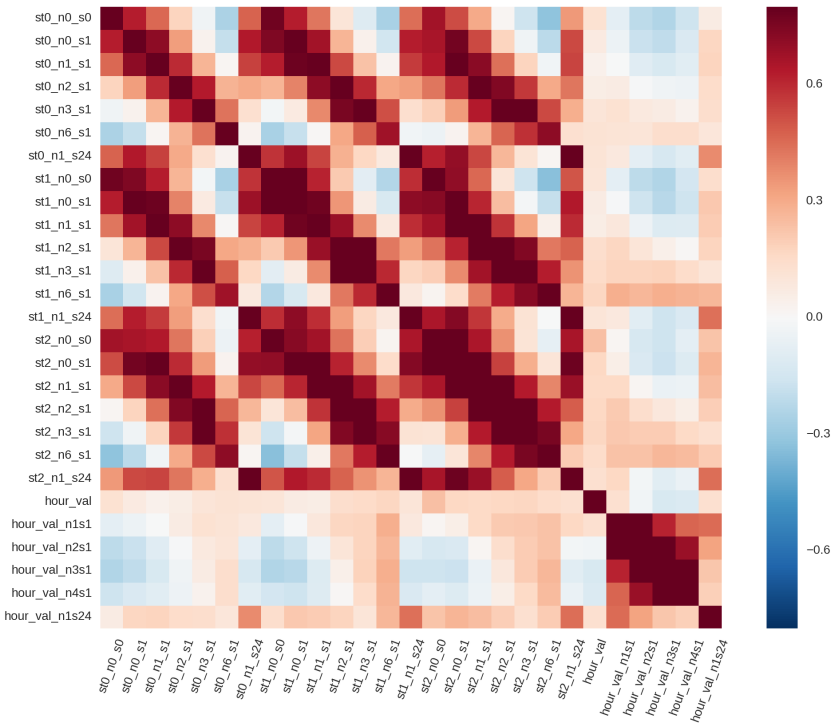


Figura E.10: Correlación entre carga de tráfico y contaminante en la estación 28079047 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.

E.0.2. Correlación entre predicciones metereológicas y concentración de NO_2

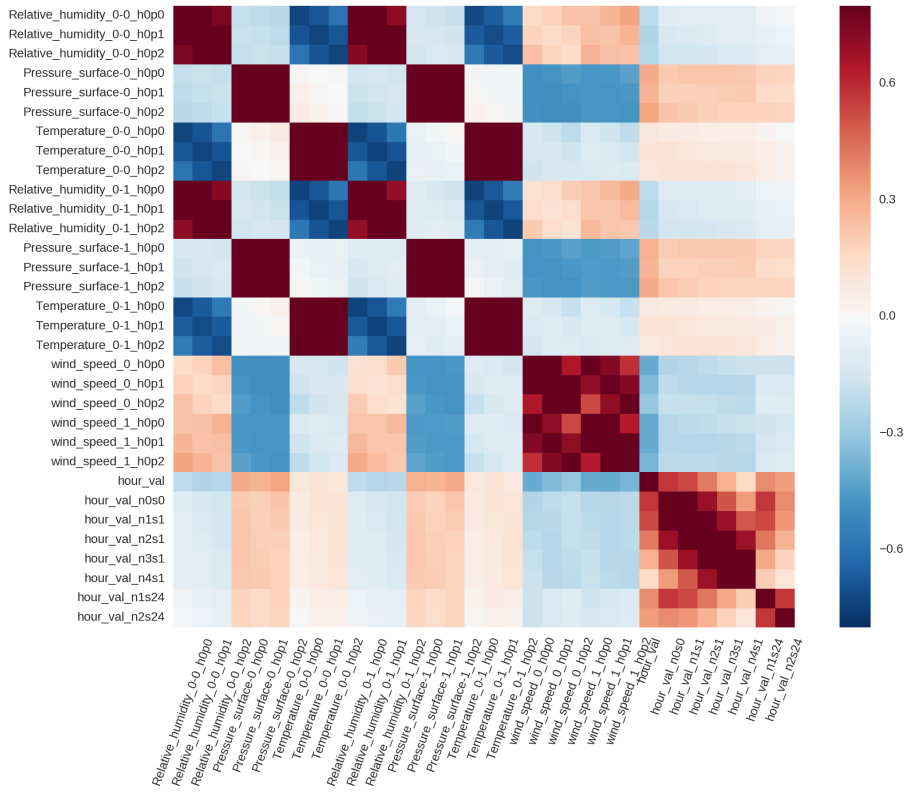


Figura E.11: Correlación entre variables metereológicas y contaminante en la estación 28079004.

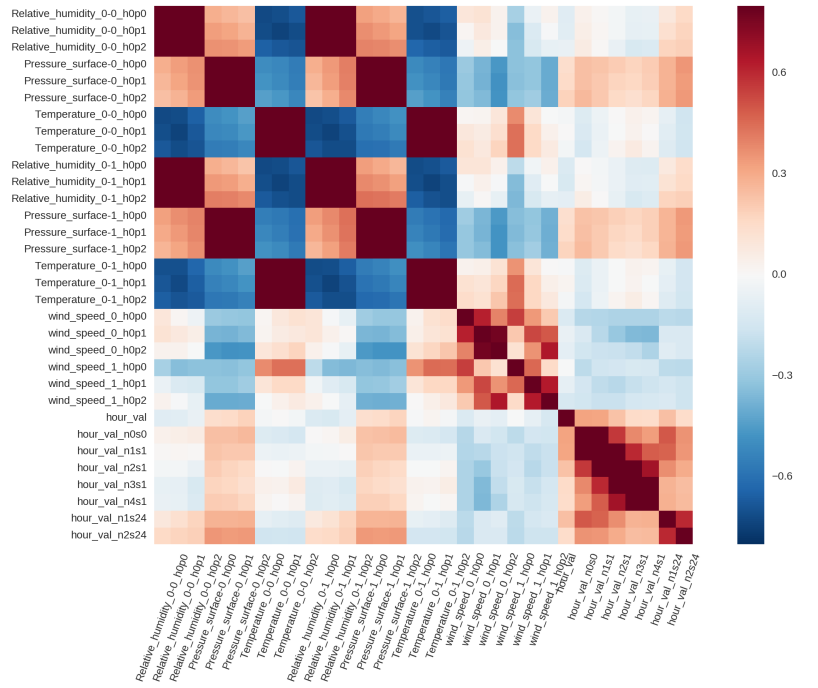


Figura E.12: Correlación entre variables metereológicas y contaminante en la estación 28079004 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.

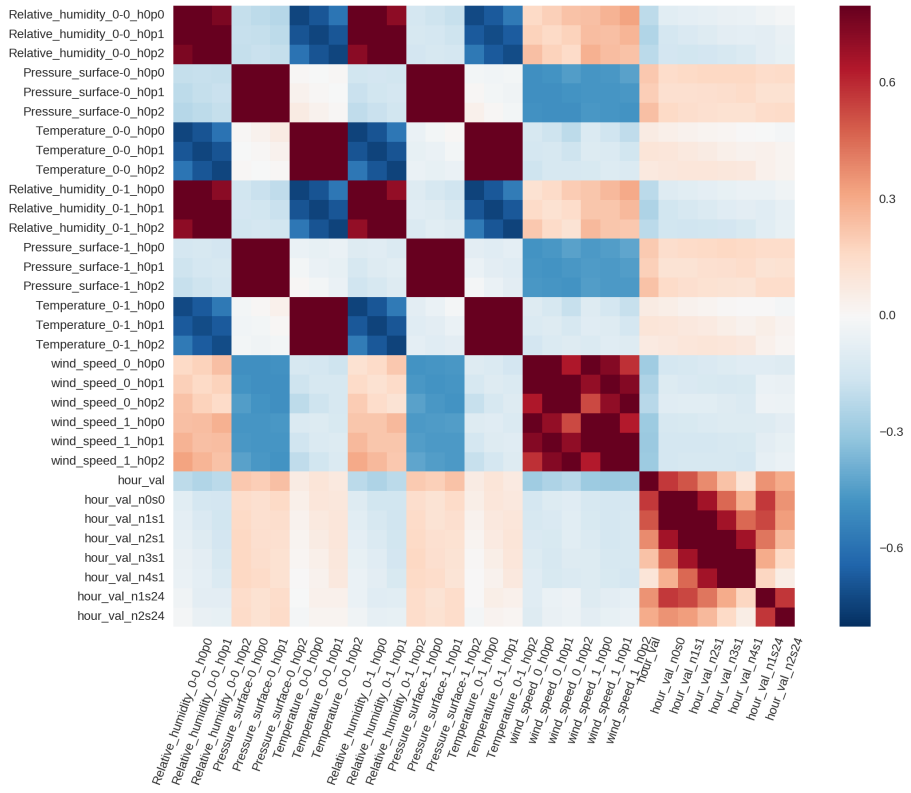


Figura E.13: Correlación entre variables metereológicas y contaminante en la estación 28079008.

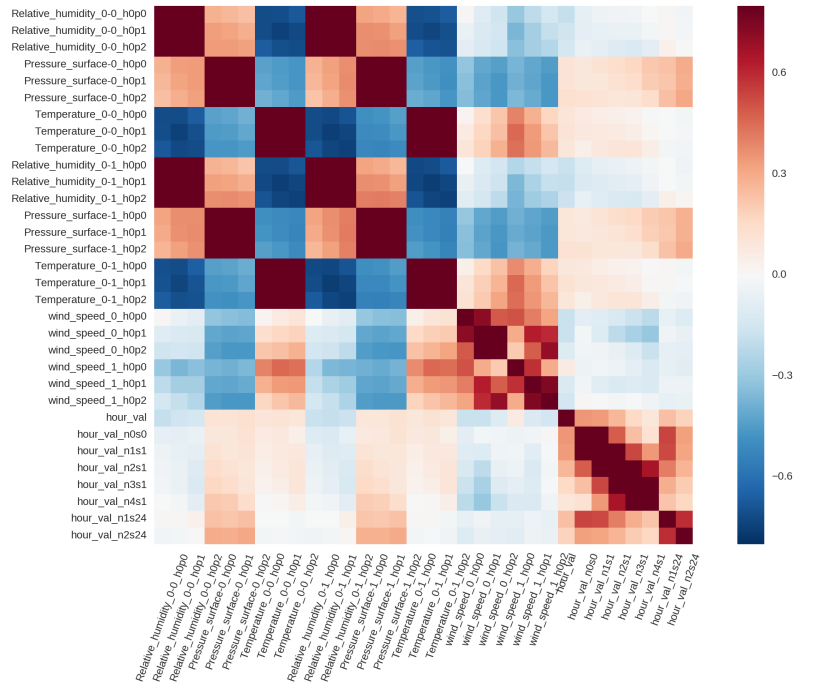


Figura E.14: Correlación entre variables metereológicas y contaminante en la estación 28079008 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.

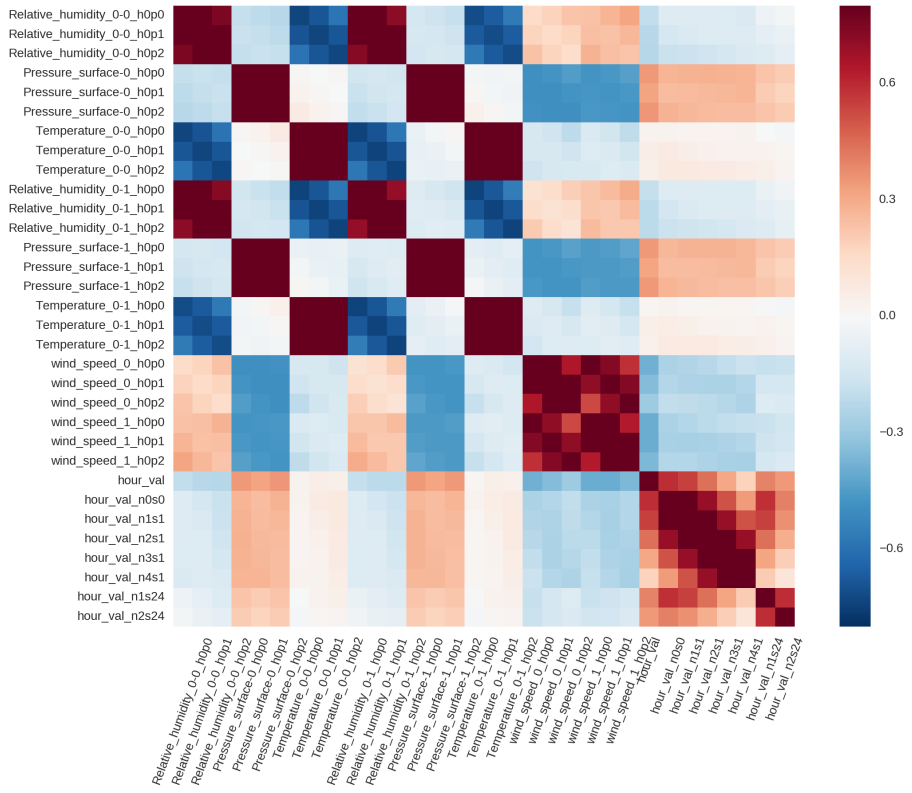


Figura E.15: Correlación entre variables metereológicas y contaminante en la estación 28079035.

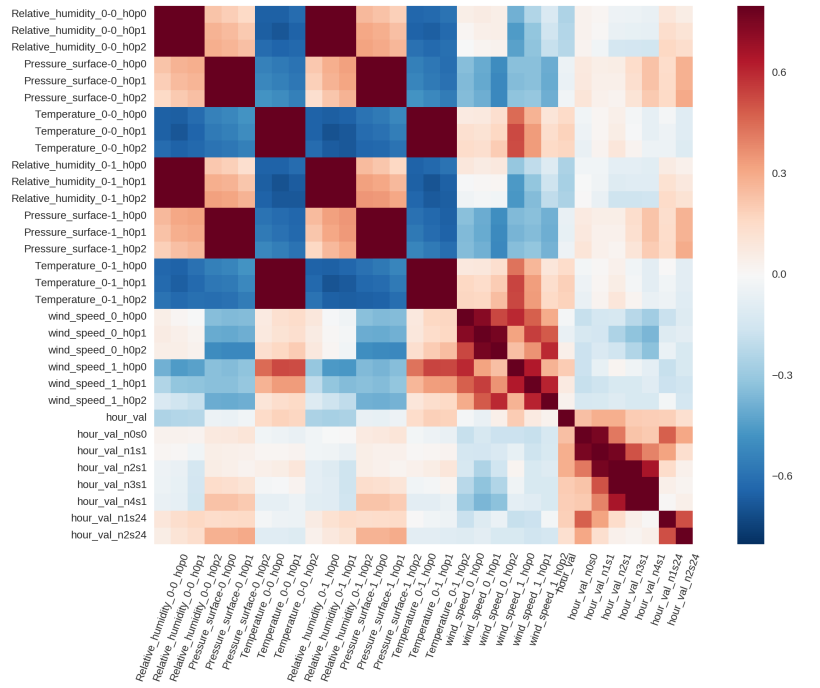


Figura E.16: Correlación entre variables metereológicas y contaminante en la estación 28079035 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.

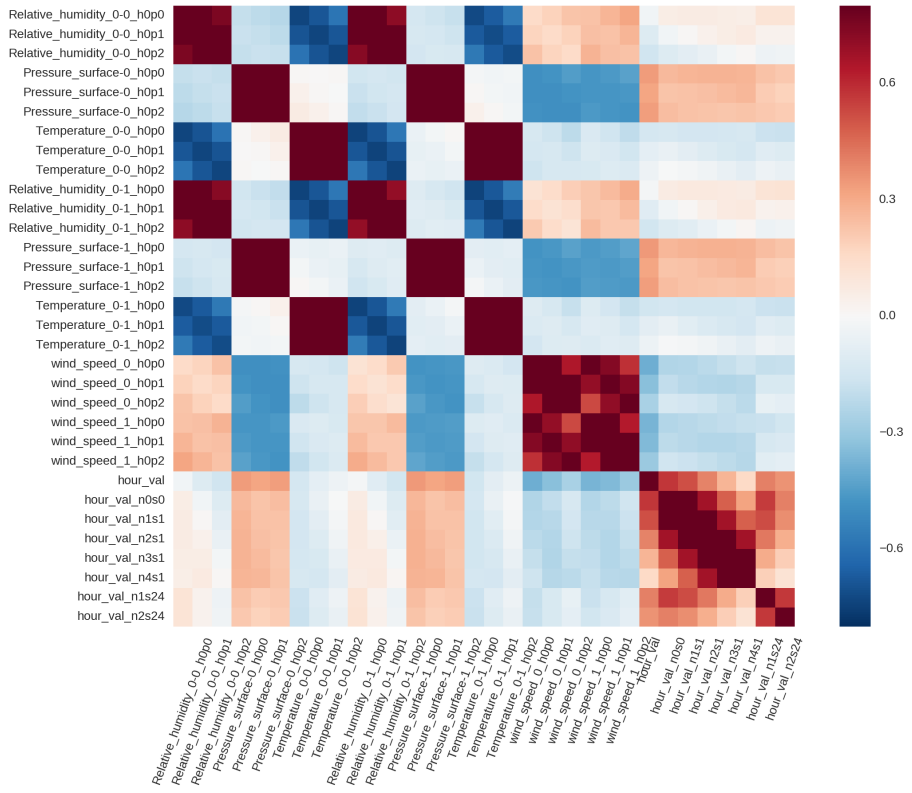


Figura E.17: Correlación entre variables metereológicas y contaminante en la estación 28079047.

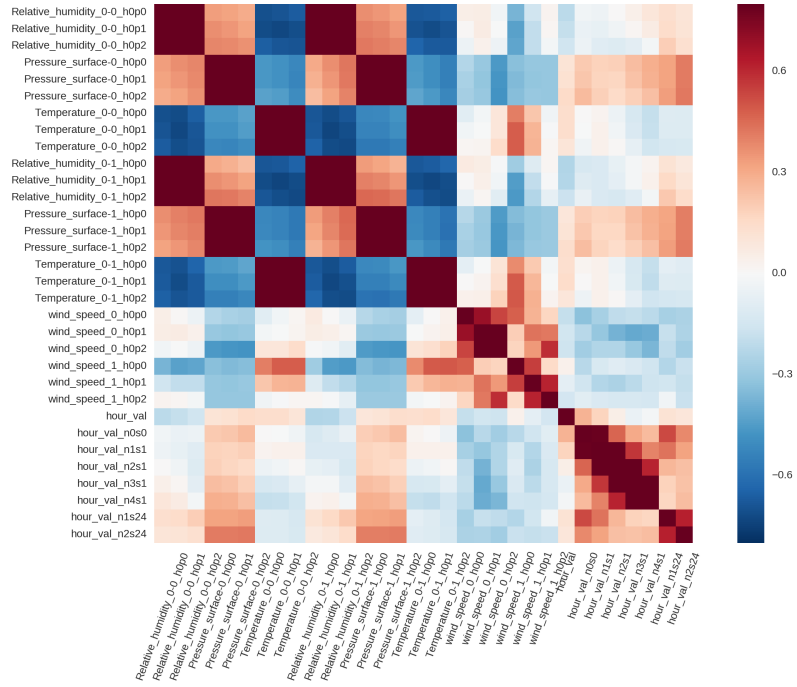


Figura E.18: Correlación entre variables metereológicas y contaminante en la estación 28079047 para valores de concentración de NO_2 mayores de 100 $\mu g/m^3$.